Adaptive Evolutionary Fusion for Multi-View Clustering

Yunxiao Zhao [1], **Liang Bai** [1] *, **Xian Yang** [2]

[1]Institute of Intelligent Information Processing, Shanxi University, Taiyuan, China
[2] Alliance Manchester Business School, The University of Manchester, Manchester, UK
202012407014@email.sxu.edu.cn, bailiang@sxu.edu.cn, xian.yang@manchester.ac.uk

Abstract

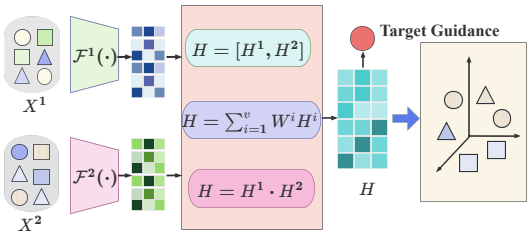
Deep multi-view clustering (MVC) methods achieve impressive performance by effectively capturing complementary information across views, where feature fusion serves as the critical mechanism for maximizing cross-view complementarity. However, most existing methods suffer from rigid dependence on non-adaptive predefined fusion operations, resulting in unverifiable and potentially suboptimal fused feature quality. To resolve these limitations, we propose a novel multi-view clustering framework that learns adaptive hierarchical fusion through an unsupervised evolutionary algorithm. Unlike conventional predefined-fusion strategies, our approach employs tree-structured representations (Fusion Trees) for adaptive feature integration. These Fusion Trees are optimized via our evolutionary mechanism, in which models sharing identical architectures but distinct Fusion Trees are conceptualized as evolutionary individuals. Through implementation of the evolutionarily optimized Fusion Tree, the resultant model generates discriminative representations in accordance with biological evolutionary principles. Comprehensive benchmarking across twelve multi-view datasets validates significant performance gains improvement over state-of-the-art baselines.

Code — https://github.com/zyxforever/AEF-MVC.git

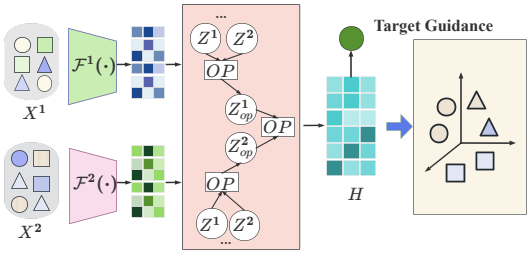
Introduction

Multi-view Clustering (MVC) techniques (Bai, Liang, and Cao 2021) are designed to leverage the complementary nature of data from multiple views to enhance clustering performance. With the explosive growth of deep learning, deep multi-view clustering has made significant progress by leveraging the powerful feature extraction capabilities of deep neural networks. These methods employ self-supervised techniques to learn view-specific representations(e.g. Multi-VAE (Xu et al. 2021)) or pursue inter-view consistency through various approaches (e.g. CVCL (Chen et al. 2023)). In terms of cross-view complementarity, they fuse features across different views to exploit complementary information.

Current multi-view clustering methods employ two distinct feature fusion paradigms with inherent limitations. The



(a) Predefined feature fusion strategy



(b) Adaptive multi-view feature fusion strategy

Figure 1: Illustration of our ideas. (a) Low-quality features generated by predefined fusion strategies will produce inferior target guidance, thereby degrading the performance of multi-view clustering. (b) The fusion strategy learned through an unsupervised evolutionary algorithms can effectively enhance the quality of fused features and generate high-quality target guidance.

early fusion approach, exemplified by DeepMVC (Trosten et al. 2023)—applies traditional clustering algorithms (K-Means (Macqueen 1967), Spectral Clustering (Shi and Malik 2000), Deep Divergence-based Clustering(Kampffmeyer et al. 2019)) to predefined fused features without integrating them into model training. Critically, this decoupled architecture does not integrate fused features into training optimization, resulting in under-utilization of complementary information and representational inefficiency. Supervision-driven frameworks (e.g., ICMVC (Chao, Jiang, and Chu 2024), DealMVC (Yang et al. 2023)) address this limitation by extracting supervisory signals (cluster assignments/similarity relations) from fused representations to guide model optimization. However, as evidenced in Fig. 1a, their depen-

*Corresponding author.

dence on predefined fusion operations (attention (Zhou and Shen 2020), concatenation (Xin, Zeng, and Wang 2021), averaging (Li et al. 2019)) critically constrains adaptability.

To address limitations in current deep multi-view clustering methods, we propose an adaptive fusion for multi-view clustering. As shown in Fig. 1b, our framework systematically explores combinatorial fusion strategies to enhance the quality of fused features, allowing for the extraction of superior target guidance for model training. We encode fusion strategies as configurable tree structures ("Fusion Trees"), where each topology defines a unique hierarchical fusion policy. Since optimal tree discovery constitutes a combinatorial optimization problem, we employ evolutionary computation following standard paradigms: each neural network equipped with a distinct Fusion Tree represents an evolutionary individual evaluated by unsupervised clustering metrics. The highest-scoring individual yields the optimal model, with its embedded Fusion Tree constituting our final solution. Experimental results demonstrate that this evolutionarily optimized framework significantly improves clustering performance through improved feature representations.

- We propose a novel deep multi-view clustering framework, which achieves discriminative feature fusion through adaptive fusion of multi-view representations.
- We represent the combination strategies of multi-view fusion via a tree structure and employ an evolutionary algorithm to discover the optimal combination in an unsupervised manner.
- Experimental results on multiple popular multi-view datasets demonstrate that our method can effectively learn fusion strategies for multi-view data features and exhibits superior performance in MVC tasks.

Related Works

Multi-view Clustering

Cross-view complementarity—essential for deep multi-view clustering—is typically achieved via feature fusion. Early approaches (e.g., DeepMVC and related works) first learn view-specific features, then employ predefined fusion strategies. Eventually, traditional clustering algorithms process these fused representations for clustering. Although initial methods underutilized fused features, recent techniques extract supervisory signals from fused representations to guide training: DealMVC (Yang et al. 2023) constructs pseudo-label and similarity graphs from concatenated features to optimize their discrepancy; Trust (Wang et al. 2023) formulates neighbor graphs from averaged features to constrain view-specific structures; GCLMVC (Xu, Yin, and Zhang 2024) and ICMVC (Chao, Jiang, and Chu 2024) generate pseudo-labels via concatenation/attention fusion for training supervision.

All of these methods are based on predefined fusion strategies, yielding suboptimal feature representations. Consequently, the extracted supervisory signals exhibit quality limitations that degrade the performance of the model. To address this fundamental constraint, we introduce adaptive view fusion, dynamically enhancing feature discriminability to significantly improve clustering efficacy.

Adaptive Fusion Strategy Solving

Since solving the adaptive view fusion strategy constitutes a combinatorial optimization problem closely tied to multimodal feature fusion in learning, the most relevant approach focuses on optimizing multimodal fusion strategies. Existing methods include those employing Gating Networks, such as DynMM (Xue and Marculescu 2023), which control fusion strategy activation but suffer from inflexible predefined combinatorial rules. Alternatively, Neural Architecture Search (NAS) techniques like MaAS (Zhang et al. 2025) and EVO (Bi et al. 2024) discover optimal network structures; for example, MFAS (Perez-Rua et al. 2019) achieves single-modal fusion exclusively through concatenation. To enable more flexible fusion, CoMo (Fu et al. 2024) leverages evolutionary algorithms, improving the accuracy in multimodal classification.

However, despite multi-view clustering's inherent unsupervised nature, current methodologies paradoxically require labeled data, severely underutilizing evolutionary computation's potential for this domain. Notably, while existing Neural Architecture Search (NAS) techniques focus exclusively on discovering optimal network structures, our approach pioneers the adaptive fusion strategy search paradigm. To bridge this gap, we propose an unsupervised evolutionary computation framework that solves our novel optimal Fusion Tree formulation, thereby resolving the adaptive view fusion challenge.

The Proposed Method

Problem Definition

Given a multi-view dataset $\{X^v = [x_1^v, \cdots, x_N^v]\}_{v=1}^V$ with N unlabeled samples, where $X^v \in R^{N \times D}$ denotes samples of the v-th view. The aim of multi-view clustering (MVC) methods is to divide these samples into K clusters.

Adaptive View Fusion Learning

Given a multi-view neural network, our aim is to learn an adaptive feature fusion strategy that hierarchically combines view-specific features to enhance representation quality. We employ tree structures ("Fusion Trees") to represent fusion strategies, where each tree defines a unique fusion methodology.

The formal definition of the Fusion Tree requires two foundational components: (i) a set of features $\mathcal{F} = \{H^1, \cdots, H^v\}$ derived from the multi-view encoder, (ii) To simplify the feature fusion process, we adopted only simplistic fusion operations, in line with previous multimodal fusion strategies (e.g., CoMo (Fu et al. 2024) and DynMM (Xue and Marculescu 2023)). a set of fusion operations $\mathcal{O} = \{Add, Mul, Concat, Max, Avg\}$.

Definition 1 ***(Leaf Node)*** *is defined as the node that contain only a feature element and possess an empty list of child nodes, expressed as follows:*

$$\mathcal{T}(f) = \langle f, \emptyset \rangle, f \in \mathcal{F} \tag{1}$$

the value of leaf nodes is the feature $val(\mathcal{T}(f)) = f$

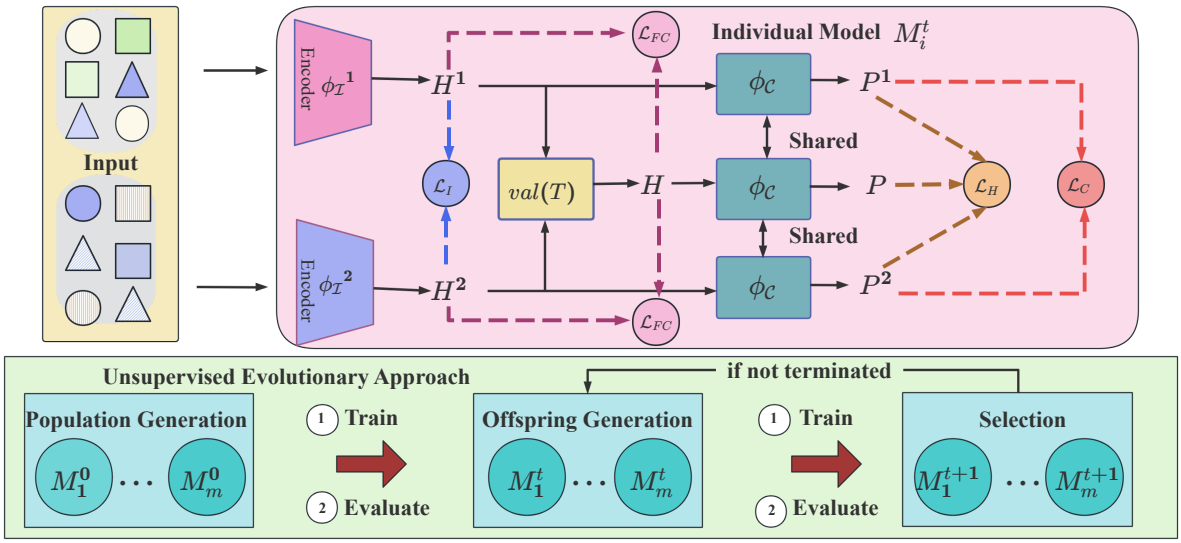


Figure 2: Framework of the proposed unsupervised evolutionary algorithm. The upper part shows individual model training, while the lower part presents the three key stages: (1) Population generation: randomly generates initial fusion trees with two leaf nodes and one operator; (2) Learning and evaluation: trains and evaluates individuals using unsupervised metrics; (3) Offspring generation: creates new populations through selection, crossover, and mutation.

Definition 2 ***(Internal Node)*** *contains an operator and possesses multiple subtrees, formally expressed as follows:*

$$\mathcal{T}(o,[T_1,\cdots,T_v]) = \langle o,[T_1,\cdots T_v]\rangle, o \in O, \\ \forall T_1,\cdots,T_v \in \mathcal{T} \tag{2}$$

where

$$\mathcal{T} = \bigcup_{f\in F}\{\langle f,\emptyset\rangle\} \cup \bigcup_{o\in O}\{\langle o,[T_1,\cdots T_v]\rangle | T_i \in \mathcal{T}\} \tag{3}$$

the value of internal node is obtained by the operator with its subtrees $val(\mathcal{T}(o,[T_1,\cdots,T_v])) = o(T_1,\cdots,T_v)$

Once an optimal fusion tree T^* has been defined, recursively traversing this tree yields the final fused feature, which is defined as:

$$H = val(T^*) \tag{4}$$

This formulation differs from existing feature fusion strategies in three key aspects: (1) The optimal tree is learnable, enabling adaptive feature fusion, unlike predefined strategies in previous work. (2) Views are adaptively selected for participation. (3) Instead of employing a single fusion strategy, our method combines multiple fusion strategies to produce the final fused feature.

At this point, we transform the combinatorial optimization problem of multi-view fusion into a tree-structured optimization problem. We will address the solution of this optimal fusion tree problem in the next section.

Evolutionary Algorithm Solving Process

This section introduces an unsupervised evolutionary approach for solving the optimal fusion tree problem. Evolutionary algorithms address optimization challenges through iterative cycles of generation, evaluation, selection, and mutation. Following this paradigm, we treat models with identical architectures but distinct tree structures as individuals, Davies-Bouldin Index (DBI) as the evaluation metric. Crossover and mutation operations on these tree structures generate new offspring, ultimately deriving the optimal fusion tree.

Unsupervised Evolutionary Approach Since the aforementioned optimization is discrete, we employ an evolutionary algorithm to tackle it. We follow the paradigm of the standard evolutionary algorithm for problem solving.

(1) Population Generation: During the initial stage, the fusion tree initially comprises only two leaf nodes (denoting features) and one root node (representing the operator), as defined as follows::

$$\mathcal{T}_{init} = \langle o_{root}, [\underbrace{\langle f_a,\emptyset\rangle}_{leaf node}, \underbrace{\langle f_b,\emptyset\rangle}_{leaf node}]\rangle \tag{5}$$

where:

$$\begin{cases} f_a, f_b \longleftarrow RandomSample(\mathcal{F},2) \\ o_{root} \longleftarrow RandomSample(\mathcal{O},1) \end{cases} \tag{6}$$

A single model $\mathcal{M}(\cdot)$ with distinct fusion trees constitutes an evolutionary individual. We formally define the initial-generation individuals as follows:

$$\mathcal{P}^0 = \{\mathcal{M}^0_{\mathcal{T}^1_{init}}(\cdot),\cdots,\mathcal{M}^0_{\mathcal{T}^m_{init}}(\cdot)\} \tag{7}$$

(2) Individual Model Evaluation: For each individual, we use the Davies-Bouldin Index (DBI) (Davies and Bouldin 1979) to assess the quality of its fused features, which serves

as the measure of its overall fitness, formally expressed as:

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\bar{D}_i + \bar{D}_j}{\|\mathbf{c}_i - \mathbf{c}_j\|}\right) \tag{8}$$

where$\bar{D}_i = \frac{1}{|C_i|}\sum_{\mathbf{x}\in C_i}\|\mathbf{x}-\mathbf{c}_i\|$ is the average intra-cluster distance for the cluster C_i, and lower values indicate superior cluster separation.

(3) Offspring Generation: After evaluating individual models, we identify the optimal individual within a generation. Subsequently, new offspring populations are generated through crossover operation, formally represented as:

$$\mathcal{Q}^{t+1} = Crossover(Selection(\mathcal{P}^t, \mathcal{Q}^t)) \tag{9}$$

Through selection, optimal models can be sequentially obtained, and we then perform pairwise ($\mathcal{M}_{\mathcal{T}_A}$, $\mathcal{M}_{\mathcal{T}_B}$) crossover between these models to generate offspring models.

$$(\mathcal{M}_{\mathcal{T}_A'}, \mathcal{M}_{\mathcal{T}_B'}) = Crossover(\mathcal{M}_{\mathcal{T}_A}, \mathcal{M}_{\mathcal{T}_B}) \tag{10}$$

where

$$\begin{cases} \mathcal{T}_A' = Replace(\mathcal{T}_A, P_A, \mathcal{S}_A) \\ \mathcal{T}_B' = Replace(\mathcal{T}_B, P_B, \mathcal{S}_B) \end{cases} \tag{11}$$

P_A and P_B are DepthWeightedSamples from $\mathcal{T}_A$ and $\mathcal{T}_B$, and $\mathcal{S}_A$ and $\mathcal{S}_B$ are subtree obtained by:

$$\begin{cases} \mathcal{S}_A = Subtree(\mathcal{T}_A, P_A) \\ \mathcal{S}_B = Subtree(\mathcal{T}_B, P_B) \end{cases} \tag{12}$$

We also mutated the newly generated tree, specifically by randomly replacing one of its subtrees with a randomly generated tree. The formula is defined as follows:

$$\mathcal{T}' = \begin{cases} Replace(\mathcal{T}, P, S_{new}) & \text{if } \xi < \rho_{rate} \\ \mathcal{T} & \text{otherwise} \end{cases} \tag{13}$$

where $\xi \sim U(0,1)$, ρ_{rate} is the mutation rate, S_{new} is a randomly generated new tree and P is DepthWeightedSamples from $\mathcal{T}$.

Through the iterative generation of new offspring and the evaluation of the resulting offspring, we can obtain the optimal model.

Individual Model Following the outline of our evolutionary algorithm for model optimization, this section details the architecture of individual models within our framework. We adopt ICMVC as the foundation, but differ from it in two key aspects: network architecture and loss functions.

(1) Network Architecture: We employ graph convolutional networks (GCN) (Kipf and Welling 2017) to extract view-specific features following the ICMVC baseline:

$$H^v = \phi_{\mathcal{I}}^v(X^v, A^v) \tag{14}$$

where A^v represents the graph structure of the v-th view. Cluster assignments are generated through:

$$\hat{Y}^v = \phi_{\mathcal{C}}^v(H^v) \in \mathbb{R}^{N\times K} \tag{15}$$

While maintaining these components from ICMVC, we replace its attention-based fusion with our adaptive multi-view fusion (Eq. 4).

(2) Loss Functions: Our approach preserves the three core objectives from ICMVC while introducing an additional regularization term. Following ICMVC, we jointly optimize: (A) $\mathcal{L}_I$: Instance-level consistency via InfoNCE (Chen et al. 2020) (B) $\mathcal{L}_C$: Cluster-level contrastive loss treating prediction columns as clusters (C) $\mathcal{L}_H$: High-confidence guidance loss generating targets from fused features

We extend this framework by introducing feature consistency regularization between pre-fusion and post-fusion representations:

$$\mathcal{L}_{FC} = \sum_{v=1}^{V}||H - H^v||_2^2 \tag{16}$$

The complete objective integrates all components:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_C + \mathcal{L}H + \mathcal{L}FC \tag{17}$$

By training models with this composite loss and evaluating individuals, we identify the optimal solution. The description of the algorithm is in the appendix.

Experiments

In this section, we primarily address the following questions: Q1 Effectiveness (I): Does the proposed AEF-MVC method outperform other methods? Q2 Visualization (I):Representative Fusion Schemes Discovered by Evolutionary Search. Q3 Clustering with Fusion (I): A Case Study of Multi-View Data. Q4 Ablation Study (I): What are the advantages of the adaptive search evolutionary multi-view fusion clustering method compared to manually predefined fusion methods? Q5:Ablation Study (II): How do key hyperparameters of the AEF-MVC framework affect its overall performance? Moreover, we conduct a more in-depth analysis of the proposed AEF-MVC through additional ablation studies, with detailed results provided in the appendix.

Experimental Settings

We use 12 datasets: **Scene-15** (Fei-Fei and Perona 2005), **LandUse-21** (Yang and Newsam 2010), **MSRC-V1** (Winn and Jojic 2005), **NoisyMNIST** (LeCun et al. 1998), **Wiki** (Rasiwasia et al. 2010), **BBCSport** (Luo et al. 2018), **BDGP** (Cai et al. 2012), and **Caltech-5V** (Xu et al. 2022). Following ICMVC's experimental configuration, we derived three datasets: **LandUse-2V** from LandUse-21 (features: GIST + LBP), **Scene-2V** from Scene-15 (features: PHOG + LBP) **MSRC-V1** from MSRC (features: GIST + HOG) Additionally, the **Caltech-2V** dataset was constructed using the first two views of Caltech-5V, maintaining feature consistency with prior work.

To evaluate the performance of the model, we adopt three evaluation metrics: Accuracy (ACC) (Chang et al. 2017), normalized mutual information (NMI) (Strehl and Ghosh 2002), and adjusted Rand index (ARI) (Hubert and Arabie 1985). Higher values for these metrics correspond to superior algorithmic performance.

Baselines Our comparative analysis includes: Post-learning fusion methods: State-of-the-art approaches to fuse features after representation learning (CoMVC (Trosten et al. 2021), AEKM (Trosten et al. 2023), DMSC (Abavisani

Datasets	Scene-2V			LandUse-2V			MSRC-V1			Noisy MNIST		
Method	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CoMVC (CVPR'2021)	37.50	38.05	20.95	26.00	32.92	13.70	60.00	52.06	39.74	36.10	26.58	17.25
AEKM (CVPR'2023)	27.58	30.04	12.76	20.95	25.91	8.22	67.14	63.08	52.33	50.43	51.41	39.43
DMSC (JSTSP'2018)	28.25	19.88	9.94	28.11	20.67	10.62	61.43	46.67	35.64	28.22	17.64	8.80
InfoDDC (CVPR'2023)	18.86	12.93	6.62	18.90	14.33	6.52	55.23	48.66	35.97	28.49	21.67	12.27
MVAE (TPAMI'2021)	19.52	20.83	5.26	26.64	22.29	10.73	79.04	65.23	58.82	66.00	60.64	50.67
MV-IIC (CVPR'2023)	24.39	21.56	13.11	22.05	18.43	10.35	53.81	55.68	42.50	47.22	59.74	45.07
SiMV (CVPR'2021)	14.71	8.79	1.98	19.67	22.34	7.20	67.72	61.80	49.80	28.43	16.67	9.58
DealMVC (ACMMM'2023)	16.43	18.35	6.61	28.14	28.14	15.91	70.95	64.64	56.07	97.30	93.18	94.17
ICMVC (AAAI'2024)	37.84	37.20	21.89	28.24	32.23	14.11	86.00	76.94	73.09	97.98	94.66	95.59
DIVIDE (AAAI'2024)	46.53	**47.26**	30.23	31.70	**39.49**	17.72	76.00	68.12	52.05	80.03	80.41	74.03
STCMCUR (AAAI'2025)	31.61	34.05	17.62	19.95	26.39	6.08	59.52	66.86	48.18	68.48	75.75	65.54
Our	**49.74**	46.39	**30.74**	**33.62**	38.23	**17.92**	**89.52**	**80.81**	**78.81**	**98.54**	**95.55**	**96.81**
Datasets	Wiki			BBCSport			BDGP			Caltech-2V		
Method	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CoMVC (CVPR'2021)	55.26	51.65	39.16	43.20	18.40	16.07	64.67	55.52	47.84	68.35	53.18	48.34
AEKM (CVPR'2023)	47.42	45.20	22.21	30.51	3.96	1.97	43.64	19.51	7.31	49.42	37.30	26.29
DMSC (JSTSP'2018)	33.39	23.87	12.58	27.72	4.45	1.73	63.64	34.39	29.51	42.92	28.03	20.69
InfoDDC (CVPR'2023)	25.71	17.18	8.18	25.00	3.20	3.95	65.08	45.88	40.54	40.14	31.00	20.99
MVAE (TPAMI'2021)	39.84	34.75	26.68	36.58	10.48	7.93	30.36	7.50	5.37	45.21	33.15	24.55
MV-IIC (CVPR'2023)	37.50	27.26	14.99	34.75	4.46	2.41	56.16	62.41	50.46	48.64	45.72	37.01
SiMV (CVPR'2021)	55.58	54.32	42.56	31.43	7.62	4.82	66.80	54.47	48.58	50.42	39.50	30.49
DealMVC (ACMMM'2023)	54.82	50.46	41.13	80.70	65.59	60.05	83.40	79.89	71.26	43.36	33.92	23.48
ICMVC (AAAI'2024)	43.93	32.96	26.72	81.62	65.76	60.36	97.80	93.32	94.61	49.76	38.04	30.32
DIVIDE (AAAI'2024)	47.73	37.89	27.73	32.17	5.42	2.42	85.27	78.57	71.73	61.57	54.88	41.04
STCMCUR (AAAI'2025)	54.74	54.99	42.58	41.18	14.85	12.33	66.08	58.18	46.05	52.28	44.17	34.66
Our	**61.10**	**57.75**	**48.49**	**87.50**	**68.81**	**73.12**	**98.84**	**96.31**	**97.13**	**73.43**	**60.65**	**55.59**

Table 1: Clustering result on 2-views datasets. The best and the second best result are denoted in bold and underline.

Datasets	Scene-15			100Leaves			Landuse-21			Caltech-5V		
Method	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CoMVC (CVPR'2021)	35.76	37.43	20.19	49.56	51.60	87.58	25.59	31.25	12.03	81.79	70.08	66.89
AEKM (CVPR'2023)	31.64	36.60	16.86	20.67	24.56	7.41	19.14	23.97	6.51	56.28	55.66	42.86
DMSC (JSTSP'2018)	30.99	24.80	13.49	69.93	83.97	56.13	25.86	26.24	10.67	54.21	37.15	29.37
InfoDDC (CVPR'2023)	24.82	24.83	10.88	20.68	53.42	7.89	19.00	20.83	7.09	68.00	60.60	54.37
MVAE (TPAMI'2021)	29.27	32.92	15.17	46.25	74.49	32.85	17.00	20.79	5.43	80.14	68.37	64.48
MV-IIC (CVPR'2023)	22.20	18.76	9.16	43.75	69.16	26.78	18.71	21.19	8.49	60.07	47.54	38.50
SiMV (CVPR'2021)	27.85	26.23	12.15	35.12	73.31	30.39	22.19	26.10	9.71	71.23	61.85	55.09
DealMVC (ACMMM'2023)	32.49	33.19	18.03	58.38	82.11	44.04	18.57	20.41	7.04	88.36	81.65	77.81
ICMVC (AAAI'2024)	39.74	37.31	22.53	64.76	86.03	55.08	25.56	30.70	12.49	68.79	58.33	51.39
DIVIDE (AAAI'2024)	46.53	47.26	30.23	76.99	88.73	68.09	32.05	**39.13**	16.85	69.14	57.34	48.05
STCMCUR (AAAI'2025)	32.33	34.95	18.17	31.06	62.13	13.29	20.57	28.06	6.43	80.85	80.35	73.44
Our	**48.49**	**47.47**	**32.34**	**78.62**	**89.44**	**69.68**	**32.43**	36.90	**17.10**	**91.36**	**83.91**	**82.09**

Table 2: Clustering result on more than two views datasets. The best and the second best result are denoted in bold and underline.

and Patel 2018), InfoDDC (Trosten et al. 2023), MVAE (Liu et al. 2021), MV-IIC (Trosten et al. 2023), SiMV (Trosten et al. 2021)). Advanced Fusion Utilization Techniques: Contemporary frameworks strategically leveraging fused features (DealMVC (Yang et al. 2023), DIVIDE (Lu et al. 2024), STCMCUR (Hu et al. 2025)).

Implementation Details. Our method is implemented in PyTorch 2.4.1 on Ubuntu 20.04 with 4 NVIDIA A40 GPUs. We use default settings for the evolutionary search: population size = 10, generations = 10, crossover rate = 0.9, and mutation rate = 0.2. For datasets with more than 3 views, generations are increased to 20 to enrich the search space. For training, we follow the ICMVC configuration: learning rate $= 0.001$, temperature parameters $\tau_I = 1.0$, $\tau_C = 0.5$, and 200 epochs. These settings provide stable and consistent performance across benchmarks.

Experimental Results

We conduct research on the aforementioned key issues by presenting detailed empirical results, thus validating the effectiveness and credibility of our model.

Q1 Effectiveness (I). To evaluate the effectiveness and robustness of the proposed method, we conduct experiments on eight representative multi-view datasets, following the evaluation metrics and settings used in previous studies. The results are summarized in Table 1. Overall, methods such

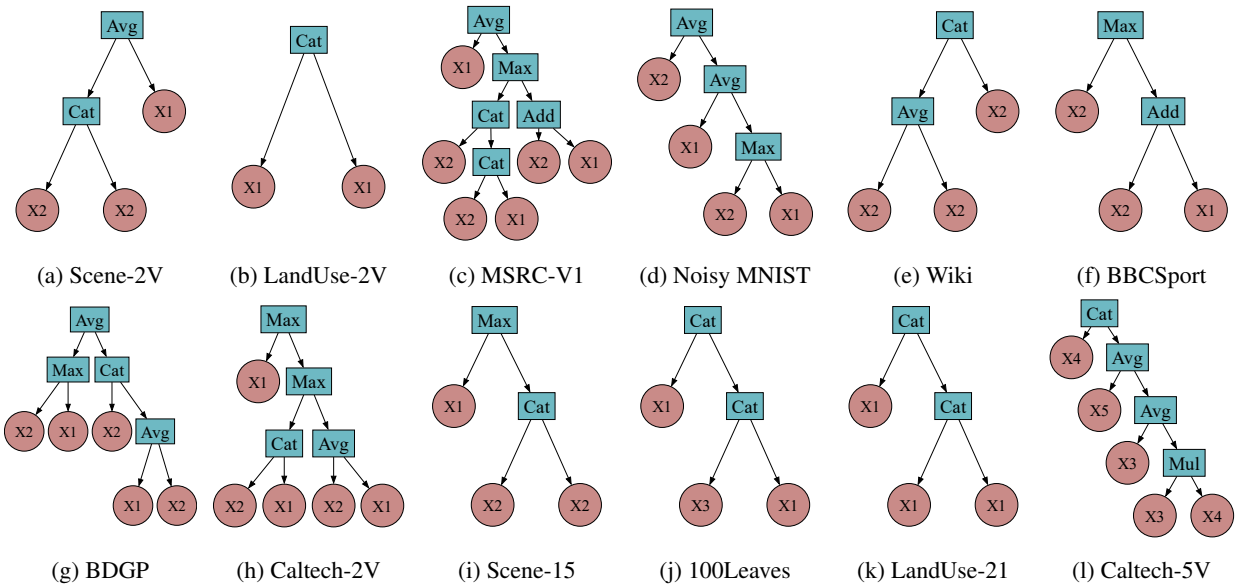


Figure 3: The final fusion tree of different datasets.

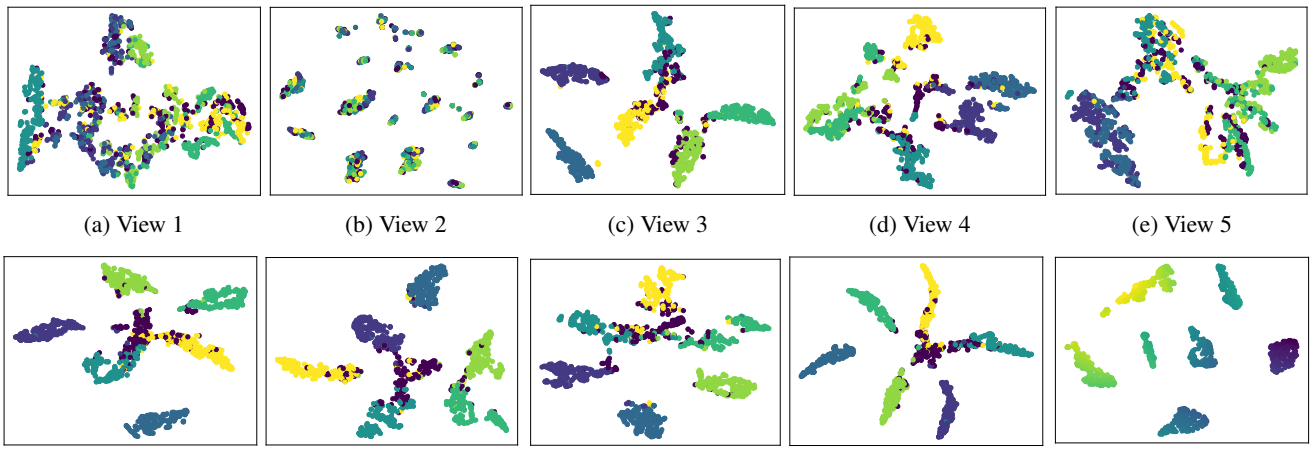
as DealMVC and ICMVC, which extract supervisory signals from fused features, outperform traditional approaches that simply concatenate view-specific features. This demonstrates the advantage of feature fusion in integrating multi-view information. However, their performance varies significantly across datasets, making it difficult to maintain consistent superiority. For example, while they perform well on Scene-2V and LandUse-2V, their accuracy drops notably on Wiki and BBCSport, mainly due to the limited adaptability of their predefined fusion strategies under varying data distributions. In contrast, the proposed AEF-MVC, equipped with an evolution-based adaptive fusion architecture, achieves more stable and superior results, showing a strong generalizability. Specifically, on Wiki and BBCSport, AEF-MVC surpasses the second-best method by 5.52% and 5.88% in ACC, respectively, further validating its effectiveness in various scenarios.

To evaluate the effectiveness of our method in complex multi-view scenarios, we conducted experiments on four representative datasets. As shown in Table 2, our method consistently achieved a top performance, demonstrating strong clustering ability and adaptability. For example, on Scene-15, our ACC reached 48.49%, outperforming ICMVC by 8.75%. On Caltech-5V, it achieved 91.36%, about 3% higher than DealMVC, with the best NMI and ARI as well. Even on LandUse-21, where view quality is uneven, our method maintained a clear advantage, highlighting its capability in view selection and robust performance. These results confirm the effectiveness and broad applicability of our adaptive fusion strategy.

Q2 Visualization. To further validate the importance of adaptive multi-view fusion strategies, we visualize the fusion structures automatically selected by the evolutionary algorithm across 12 datasets. As shown in Figure 3, the selected structures differ significantly in terms of operator types, fusion tree depth, and path patterns, reflecting dataset-specific characteristics in multi-view integration. This confirms the need for fusion strategies to be adaptively adjusted based on data properties. For example, on the LandUse-21 dataset, the fusion structure consistently relies only on the first view X_1, omitting the second view—indicating that the latter may be redundant or noisy, offering limited complementary information. These observations demonstrate that our method not only dynamically adjusts fusion operations but also automatically identifies informative views, thereby mitigating the negative impact of low-quality views on overall performance.

Q3 Clustering with Fusion: Figure 4 shows clustering visualizations on a five-view dataset using different methods. Subfigures (a–e) present the results from individual views, (f–h) show improved clustering on the adaptively selected views (Views 3–5), (i) is the final fused result from AEF-MVC, and (j) is the result from the baseline STCMCUR. As observed, Views 1 and 2 yield scattered and overlapping clusters, while Views 3–5 perform better but still suffer from intra-class and inter-class issues. AEF-MVC effectively selects informative views (3–5) and discards noisy ones (1–2), enabling better feature integration. The improved single-view results (f–h) highlight enhanced representations after adaptive fusion. The final result (i) forms more compact and well-separated clusters, clearly outperforming STCMCUR (j), and validating our method's effectiveness in both view selection and representation learning.

Q4 Ablation Study (I). To evaluate the effectiveness of the proposed AEF-MVC method, we conducted experiments on four multi-view datasets. We compared AEF-MVC against five representative manually predefined fusion strategies: Add, Mul, Cat, Max, and Avg As shown in Figure 5,the experimental results show that AEF-MVC consistently achieved the best clustering performance across all



(a) View 1 (b) View 2 (c) View 3 (d) View 4 (e) View 5

(f) Improved view 3 (g) Improved view 4 (h) Improved view 5 (i) Fused View AEF-MVC (j) Fused View STCMCUR

Figure 4: t-SNE visualizations of five individual views and five fusion strategies on Caltech-5V.

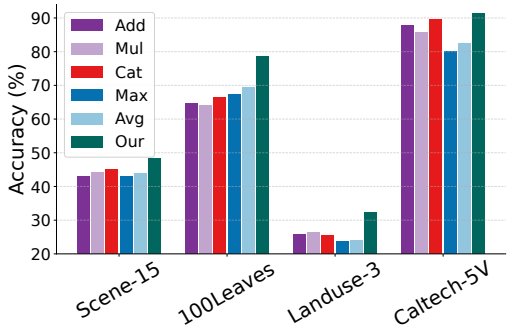


Figure 5: Clustering accuracy comparison of AEF-MVC and manual fusion methods on four multi-view datasets.

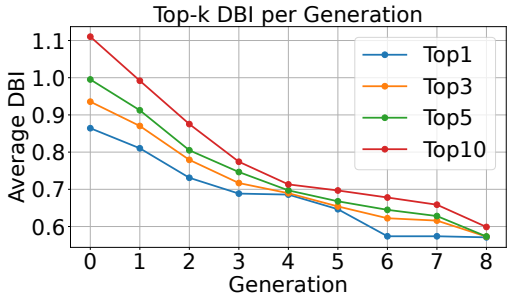


Figure 6: Top-k DBI Performance Across Generations (Lower is Better), on Caltech-5V datasets.

datasets, demonstrating strong fusion capability and robustness. In particular, on more complex datasets such as Scene-15 and Landuse-21, AEF-MVC outperformed the best baseline methods by approximately 3.4% and 6.1%, respectively. Even in high-performing datasets like Caltech-5V, it further improved clustering accuracy. Unlike manually defined fusion schemes based on empirical heuristics, AEF-MVC automatically optimizes fusion structures through evolutionary search, offering superior adaptability and generalization. These results validate the effectiveness and superiority of AEF-MVC.

Q5 Ablation Study (II). This experiment evaluates the DBI performance of the Top 1, Top 3, Top 5, and Top 10 individuals across generations. As shown in Figure 6, lower DBI values indicate better clustering quality.The results reveal a clear downward trend in DBI for all Top-k curves during the early generations, indicating effective optimization of clustering performance. Notably, the Top 1 line shows the sharpest decline, followed by Top 3 and Top 5. After generation 6, all curves tend to stabilize, suggesting that the algorithm has largely converged. This implies that high-quality individuals have become dominant in the population, and further iterations bring only marginal improvements. Overall, these findings confirm that the proposed method exhibits good convergence behavior and stability, and can consistently discover high-quality clustering structures through evolutionary optimization.

Conclusion

This paper presents a novel multi-view clustering framework that systematically investigates adaptive hierarchical feature fusion strategies across diverse views, formally establishing the problem as an evolutionary-optimized Fusion Tree construction task. Through an unsupervised evolutionary optimization framework that incorporates genetic operators, we effectively derive discriminative feature learning. Comprehensive benchmark evaluations on several datasets demonstrate the superior efficacy of our method. In future, leveraging the inherent characteristics of our algorithm, we will explore extensions to incomplete multi-view data scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62432006, 62276159) and the Fundamental Research Program of Shanxi Province (No.202303021223004).

References

Abavisani, M.; and Patel, V. M. 2018. Deep Multimodal Subspace Clustering Networks. *IEEE Journal of Selected Topics in Signal Processin*, 12(6): 1601–1614.

Bai, L.; Liang, J.; and Cao, F. 2021. Semi-Supervised Clustering With Constraints of Different Types From Multiple Information Sources. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9): 3247–3258.

Bi, Y.; Liang, J. J.; Xue, B.; and Zhang, M. 2024. A Genetic Programming Approach With Building Block Evolving and Reusing to Image Classification. *IEEE Transactions on Evolutionary Computation*, 28(5): 1366–1380.

Cai, X.; Wang, H.; Huang, H.; and Ding, C. 2012. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12): 116–124.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep Adaptive Image Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5880–5888.

Chao, G.; Jiang, Y.; and Chu, D. 2024. Incomplete Contrastive Multi-View Clustering with High-Confidence Guiding. In *AAAI Conference on Artificial Intelligence*, 11221–11229.

Chen, J.; Mao, H.; Woo, W. L.; and Peng, X. 2023. Deep Multiview Clustering by Contrasting Cluster Assignments. In *IEEE International Conference on Computer Vision*, 16706–16715.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*.

Davies, D. L.; and Bouldin, D. W. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2): 224–227.

Fei-Fei, L.; and Perona, P. 2005. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fu, P.; Liang, X.; Qian, Y.; Guo, Q.; Wei, Z.; and Li, W. 2024. CoMO-NAS: Core-Structures-Guided Multi-Objective Neural Architecture Search for Multi-Modal Classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9126–9135.

Hu, S.; Tian, B.; Liu, W.; and Ye, Y. 2025. Self-supervised Trusted Contrastive Multi-view Clustering with Uncertainty Refined. In *AAAI Conference on Artificial Intelligence*, 17305–17313.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Kampffmeyer, M.; Løkse, S.; Bianchi, F. M.; Livi, L.; Salberg, A.; and Jenssen, R. 2019. Deep divergence-based approach to clustering. *Neural Networks*, 113: 91–101.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11): 2278–2324.

Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; and Yang, Z. 2019. Deep Adversarial Multi-view Clustering Network. In *International Joint Conference on Artificial Intelligence*, 2952–2958.

Liu, X.; Liu, L.; Liao, Q.; Wang, S.; Zhang, Y.; Tu, W.; Tang, C.; Liu, J.; and Zhu, E. 2021. Multi-VAE: Learning Disentangled View-common and View-peculiar Visual Representations for Multi-view Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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 *AAAI Conference on Artificial Intelligence*, 14193–14201.

Luo, S.; Zhang, C.; Zhang, W.; and Cao, X. 2018. Consistent and Specific Multi-View Subspace Clustering. In *AAAI Conference on Artificial Intelligence*, 3730–3737.

Macqueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

Perez-Rua, J. M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. MFAS: Multimodal Fusion Architecture Search. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Rasiwasia, N.; Pereira, J. C.; Coviello, E.; Doyle, G.; Lanckriet, G. R. G.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia*, 251–260.

Shi, J.; and Malik, J. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888–905.

Strehl, A.; and Ghosh, J. 2002. Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3: 583–617.

Trosten, D. J.; Løkse, S.; Jenssen, R.; and Kampffmeyer, M. 2021. Reconsidering Representation Alignment for Multi-View Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1255–1265.

Trosten, D. J.; Løkse, S.; Jenssen, R.; and Kampffmeyer, M. C. 2023. On the Effects of Self-supervision and Contrastive Alignment in Deep Multi-view Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 23976–23985.

Wang, J.; Feng, S.; Lyu, G.; and Gu, Z. 2023. Triple-Granularity Contrastive Learning for Deep Multi-View Subspace Clustering. In *ACM International Conference on Multimedia*, 2994–3002.

Winn, J. M.; and Jojic, N. 2005. LOCUS: Learning Object Classes with Unsupervised Segmentation. In *IEEE International Conference on Computer Vision*, 756–763.

Xin, B.; Zeng, S.; and Wang, X. 2021. Self-Supervised Deep Correlational Multi-View Clustering. In *International Joint Conference on Neural Networks*, 1–8. IEEE.

Xu, B.; Yin, J.; and Zhang, N. 2024. Graph based Consistency Learning for Contrastive Multi-View Clustering. In *ACM International Conference on Multimedia*, 8633–8641.

Xu, J.; Ren, Y.; Tang, H.; Pu, X.; Zhu, X.; Zeng, M.; and He, L. 2021. Multi-VAE: Learning Disentangled View-common and View-peculiar Visual Representations for Multi-view Clustering. In *IEEE International Conference on Computer Vision*, 9214–9223.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level Feature Learning for Contrastive Multi-view Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16030–16039.

Xue, Z.; and Marculescu, R. 2023. Dynamic Multimodal Fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2575–2584. IEEE.

Yang, X.; Jin, J.; Wang, S.; Liang, K.; Liu, Y.; Wen, Y.; Liu, S.; Zhou, S.; Liu, X.; and Zhu, E. 2023. DealMVC: Dual Contrastive Calibration for Multi-view Clustering. In *ACM International Conference on Multimedia*, 337–346.

Yang, Y.; and Newsam, S. D. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, 270–279.

Zhang, G.; Niu, L.; Fang, J.; Wang, K.; Bai, L.; and Wang, X. 2025. Multi-agent Architecture Search via Agentic Supernet. In *International Conference on Machine Learning*.

Zhou, R.; and Shen, Y. 2020. End-to-End Adversarial-Attention Network for Multi-Modal Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 14607–14616.