Diversity-oriented Deep Multi-modal Clustering

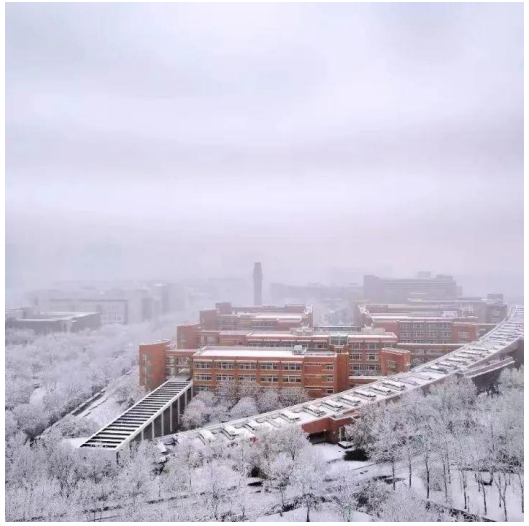
Yanzheng Wang[#], Xin Yang[#], Yujun Wang, Shizhe Hu[*], Mingliang Xu[*]

School of Computer and Artificial Intelligence

Zhengzhou University

Zhengzhou, Henan, China

Zhengzhou University (Also called “Western Park of Zhengzhou”)



Tourist Spot

Outline

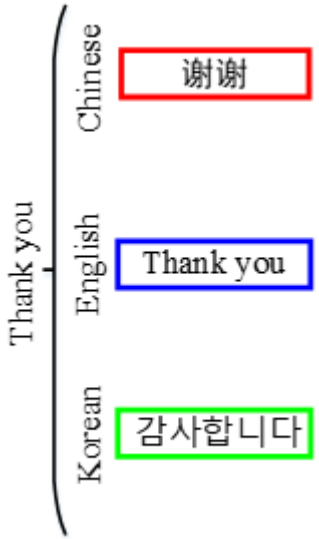
- Problem background
- Previous works
- Our proposal
- Experiments
- Conclusion

Outline

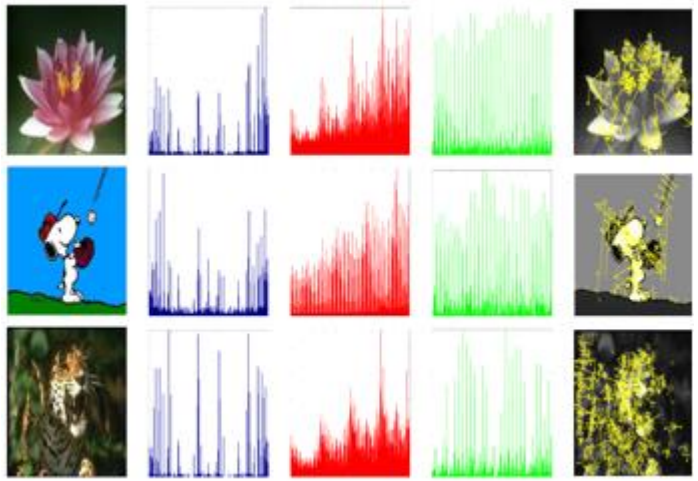
- **Problem background**
- Previous works
- Our proposal
- Experiments
- Conclusion

Characteristics of multi-modal datasets

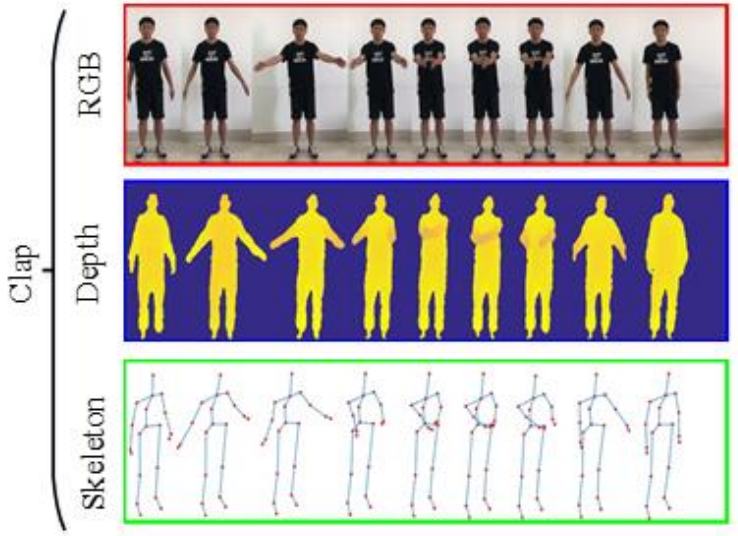
In Big Data era, many kinds of multi-modal data are emerging.



Multi-lingual Text



Multi-feature Image

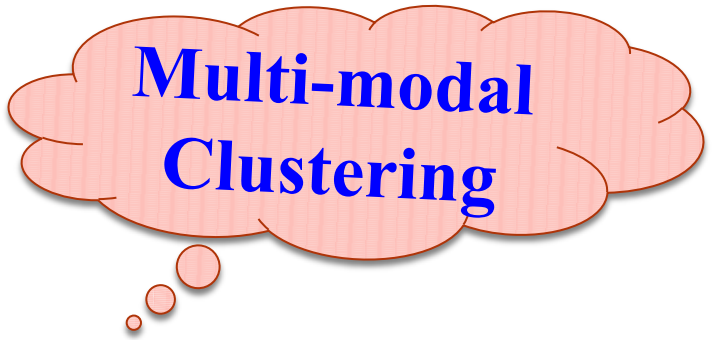


Multi-modal human action video

Property: Heterogeneous, Large-scale, Diversification, Complexity

Limitations of supervised multi-modal classification methods

1. Time-consuming and cost-expensive for labelling;
2. Over-reliance on the label information of trained data;
3. Ignoring the characteristics of the input data itself.



Challenges of existing multi-modal clustering methods

Most existing methods attempt to investigate the consistency or/and complementarity information by fusing all modalities, but this will lead to the following challenges:

1. **Information conflicts between modalities emerge.;**
2. **Information-rich modalities may be weakened**

Outline

- Problem background
- **Previous works**
- Our proposal
- Experiments
- Conclusion

Previous multi-modal clustering methods

- **Traditional multi-modal clustering methods :**

Existing traditional MMC methods mainly focus on three categories: subspace learning, graphical models and matrix decomposition (Cai et al., 2011; Xia et al., 2023).

1. Cai, X., Nie, F., Huang, H., and Kamangar, F. Heteroge neous image feature integration via multi-modal spectral clustering. In CVPR, pp. 1977–1984, 2011.

2. Xia, W., Wang, T., Gao, Q., Yang, M., and Gao, X. Graph embedding contrastive multi-modal representation learning for clustering. IEEE TIP, 32:1170–1183, 2023.

Previous multi-modal clustering methods

- **Deep multi-modal clustering methods :**

Zhou and Shen (Zhou and Shen., 2020) propose a method that first use an adversarial regularizer to align modalities, and then perform an attention fusion on all modalities, so as to quantify the importance of different modalities.

Xu et al. (Xu et al., 2022)design a feature learning model with multiple levels that learns more discriminative features from the feature/cluster-level contrast for clustering data of multiple views.

Morales et al. (Morales et al., 2018) train a separate model for each modality, then get the predictions for each modality, and finally train a new model on these new vectors to output the final prediction.

Limitations:

– *Ignore the performance degradation caused by modality fusion.*

Outline

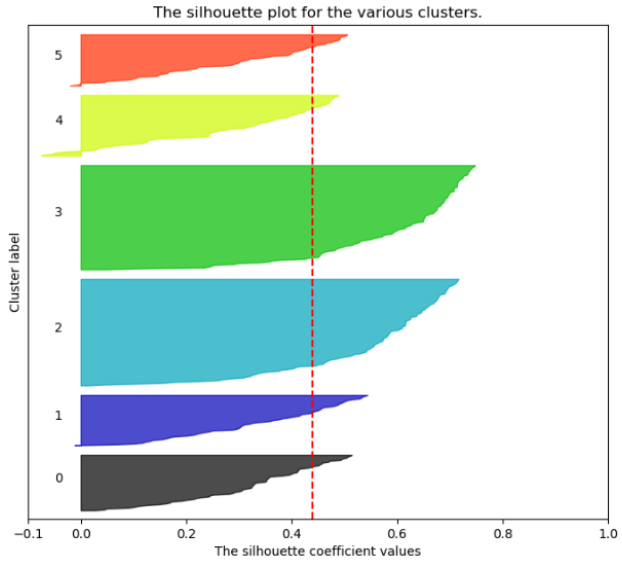
- Problem background
- Previous works
- **Our proposal**
- Experiments
- Conclusion

Our proposed method

- Diversity-oriented Deep Multi-modal Clustering (DDMC):
  - Selection of Dominant Modality Part;
  - Diversity Learning Part;
  - Clustering Part.

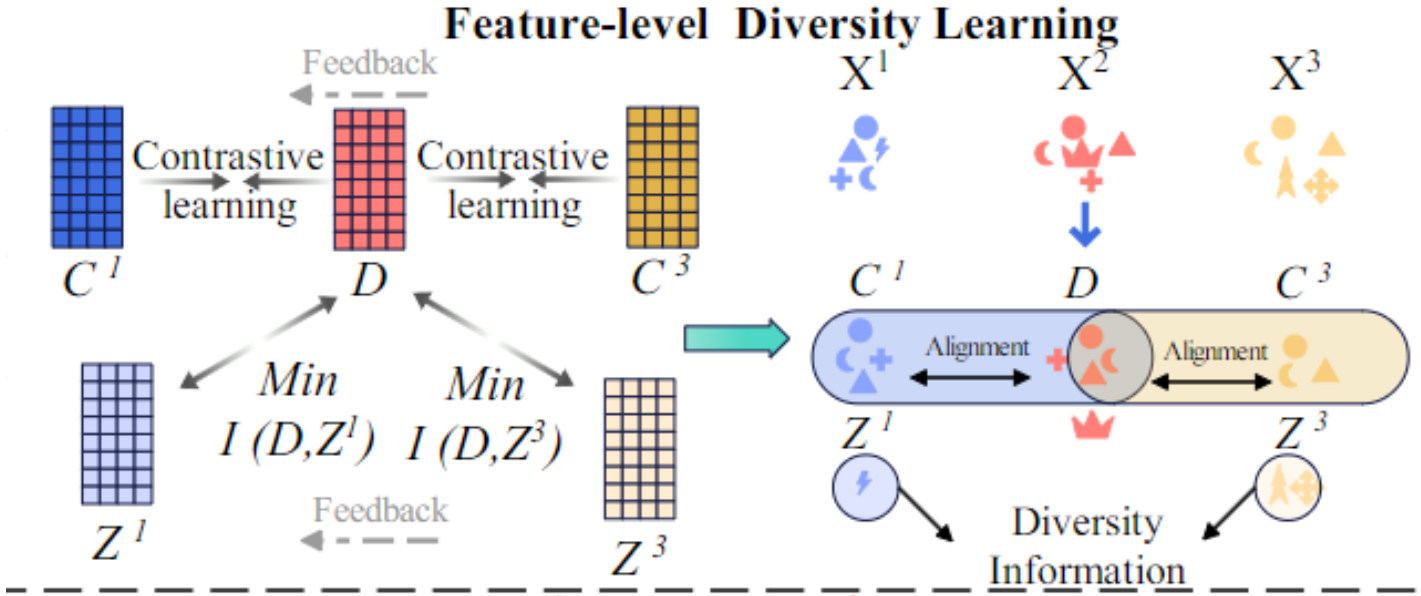
Selection of Dominant Modality Part

We decide that the dominant modality method is prior knowledge or the Silhouette coefficient (SI) fraction. Prior knowledge is that the modality weights in the dataset have been measured in previous work, or the importance of certain modality in some datasets is obvious. In the absence of reliable prior information, the SI fraction is adopted to determine the dominant modality.SI is a metric used to evaluate the quality of clustering results. It combines the intra-cluster closeness and inter-cluster separation, and calculates a SI for each data point to measure its similarity to its own cluster and its nearest neighbor cluster.



Diversity Learning Part

In the feature diversity learning level, feature diversity learning is achieved through information compression and contrastive learning.



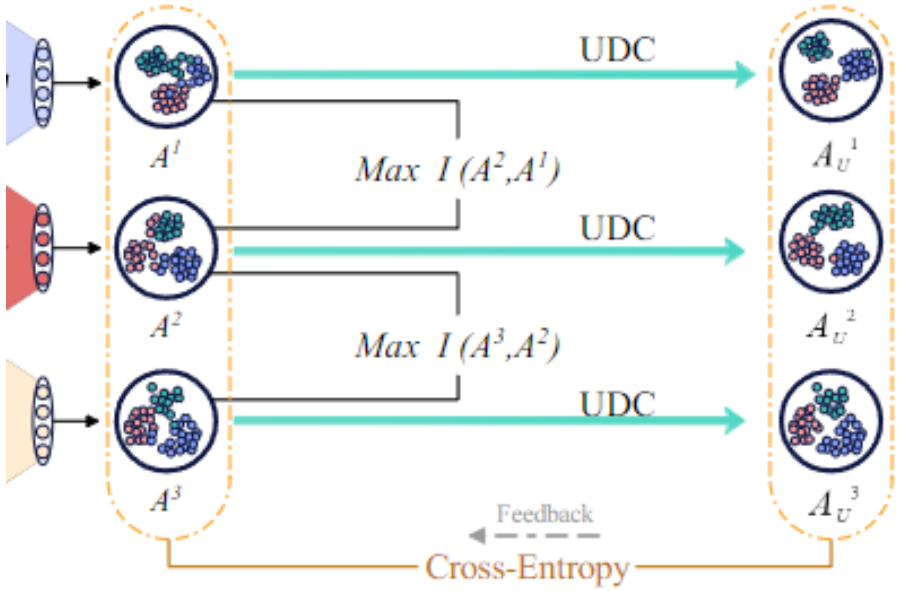
$$\mathcal{L}_1 = I(X^{dm}; D^{dm}) + \sum_{m \neq dm}^{M} I(X^m; Z^m) + \sum_{m \neq dm}^{M} I(Z^m; D^{dm}).$$

$$\mathcal{L}_2 = \frac{1}{2}\left(\sum_{m \neq dm}^{M} \ell^{dm,m} + \sum_{m \neq dm}^{M} \ell^{m,dm} \right)$$

$$\mathcal{L}_{FDL} = \mathcal{L}_1 + \mathcal{L}_2.$$

Diversity Learning Part

In cluster-level contrastive learning, cluster-level feature extraction is achieved through Uniform Distribution Constraint and mutual information methods.



$$\mathcal{L}_3 = \sum_{m \neq dm}^{M} I(A^m; A^{dm}) \qquad \mathcal{L}_4 = \sum_{m}^{M} CE(A^m; {A_U}^m) \qquad \mathcal{L}_{CDL} = \mathcal{L}_4 - \mathcal{L}_3.$$

Clustering Part

The enhanced dominant modality is clustered by Deep Divergence-based Clustering to obtain the cluster assignment matrix Q. DDC is an effective unsupervised clustering method in deep learning, which aims to improve clustering performance by optimizing the dissimilarity measure between sample distributions. It consists of three parts, the first part is intra-cluster compression and inter-cluster separation, the second part is the orthogonality constraint, and the third part is the assignment of simple simplex corners, the loss optimization function of the clustering module is as follows:

$$\mathcal{L}_{DDC} = \frac{1}{K}\sum_{i=1}^{K-1}\sum_{j>i}\frac{\delta_i^T \mathbf{K}\delta_j}{\sqrt{\delta_i^T \mathbf{K}\delta_i \delta_j^T \mathbf{K}\delta_j}} + \mathrm{triu}(Q^T Q) + \frac{1}{K}\sum_{i=1}^{K-1}\sum_{j>i}\frac{\lambda_i^T \mathbf{K}\lambda_j}{\sqrt{\lambda_i^T \mathbf{K}\lambda_i \cdot \lambda_j^T \mathbf{K}\lambda_j}}.$$

Objective function

We propose a novel loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{DDC} + \alpha \mathcal{L}_{FDL} + \beta \mathcal{L}_{CDL}.$$

α、β are trade-off parameters used to balance feature-level and cluster-level diversity learning.

Advantages of the DDMC

- Ours is the first work to investigate multi-modal clustering by enhencing dominant modalities rather than fusing modalities.

- We propose a diversity-oriented deep multi-modal clustering method by dominant modality enhancement rather than modality fusion, which can maximally retain important information in the raw modality and has the advantages of both single-modal and multi-modal clustering.

Outline

- Problem background
- Previous works
- Our proposal
- **Experiments**
- Conclusion

Datasets

Dataset	Samples	Clusters	Dimension
Caltech-3V	1440	7	40/254/928
Caltech-4V	1440	7	40/254/928/512
ESP-Game	11032	7	300/300/300
Flickr	12154	6	100/100/100
IAPR	7855	6	100/100

Compared methods

1) **Single-Modal Clustering**: K-Means (KM) and Normalized Cuts (Ncuts).
2) **All-Modal Clustering**: AmKM and AmNcuts.
3) **Traditional Clustering**:

(1) CoregMVSC: A multi-modal spectral clustering method that applies co-regularization to the clustering results.

(2) RMKMC: A multi-modal k-means clustering method that adaptively adjusts modality weights.

(3) SwMC: A totally self-weighted multi-modal clustering method for automatic modality weighting.

(4) SMCMB: This method mining rich information in multi-view data by joint learning of multiple bipartite graphs, and maintaining high efficiency on large-scale data sets, the time and space complexity is close to linear.

(5)ONMMSC:This method mining rich information in multi-view data by joint learning of multiple bipartite graphs, and maintaining high efficiency on large-scale data sets, the time and space complexity is close to linear.

Compared methods

4) Deep Clustering:

(1) EAMC:An end-to-end adversarial attention network that aligns potential feature distributions and quantifies modal importance, respectively, through adversarial learning and attention mechanisms.

(2) DEMVC:A multi-view clustering algorithm that utilizes common and complementary information from multiple views to achieve better clustering performance through deep embedded representation learning and collaborative training mechanisms.

(3) SiMVC and CoMVC: SiMVC is a simple baseline model for deep clustering. CoMVC builds on this by introducing a contrastive alignment module to overcome the limitations of traditional alignment methods.

(4) MFLVC: A hierarchical feature learning clustering method that efficiently integrates multi-level feature learning and contrastive learning.

(5) DIVIDE: A novel robust multi-view clustering method, which identifies global data pairs via high-order random walks and employs a decoupled contrastive learning framework to perform intra-view and inter-view contrastive learning in separate embedding spaces, thereby enhancing clustering performance and robustness against missing views.

(6) ICMVC: An end-to-end clustering method that handles missing data through multi-modal consistency transfer and graph convolutional networks, and combines contrastive learning.

(7) DIVIDE: A multi-modal clustering method based on decoupled contrastive learning and high-order random walks, and integrates the idea of contrastive learning to improve clustering performance.

(8) SSLNMVC:A deep multi-view clustering method that enhances the consistency of multi-view features through a consensus high-level feature learning module and aligns view-specific and view-consensus semantic labels using a self-supervised semantic calibration module.

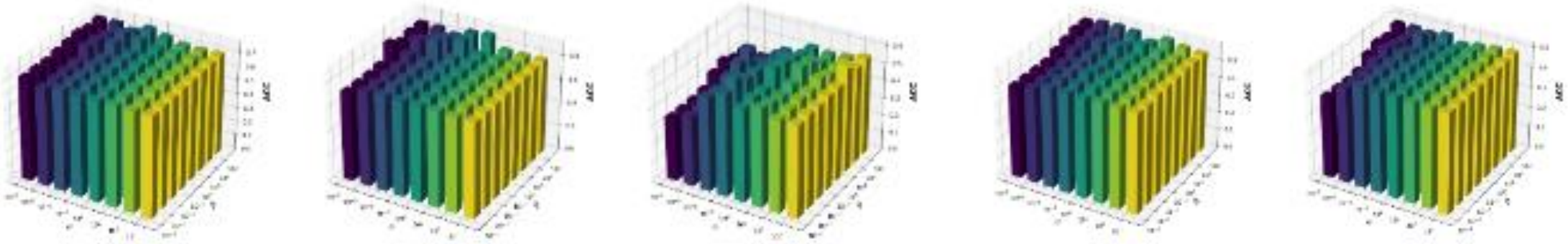
(9) SEM:This method solves the representation degradation problem caused by contrast learning in multi-view scenarios through self-weighting and information reconstruction strategies.

(10) SCMVC:The method establishes a hierarchical feature fusion framework and a self-weighted contrastive fusion approach, effectively separating the consistency objective from the reconstruction objective.

Clustering results

Methods	Caltech-3V		Caltech-4V		ESP-Game		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	46.3	31.3	54.6	46.7	43.2	29.4	40.9	22.5	38.9	17.2
Ncuts(TPAMI'00)	42.6	25.4	67.8	47.6	41.0	25.9	48.4	26.1	41.9	18.9
AmKM	46.9	31.5	44.9	30.6	49.9	34.7	41.0	21.6	40.4	17.0
AmNcuts(TPAMI'00)	43.7	25.5	41.8	24.9	33.5	19.1	48.2	26.2	42.2	18.9
CoregMVSC (NeurIPS'11)	54.4	45.3	64.9	54.5	40.1	28.8	41.0	26.8	35.1	18.4
RMKMC (IJCAI'13)	59.5	49.4	65.5	60.3	44.7	29.7	42.3	23.4	36.4	15.9
SwMC (IJCAI'17)	30.2	23.1	43.7	44.2	43.7	44.2	34.3	34.5	30.2	23.1
ONMSC (AAAI'20)	58.2	56.8	62.3	66.1	17.1	18.1	30.6	16.4	21.6	11.1
SMCMB (TBD'23)	67.2	54.5	74.4	67.0	54.9	40.5	52.8	32.1	34.8	16.4
EAMC (CVPR'20)	38.9	21.4	29.6	16.5	27.1	6.5	30.5	9.1	37.1	16.4
DEMVC (InfoSci'21)	38.7	27.0	48.4	39.7	35.5	21.6	44.8	25.2	30.1	13.8
SiMVC (CVPR'21)	56.9	50.4	61.9	53.6	35.3	16.2	45.6	26.3	42.7	18.5
CoMVC (CVPR'21)	54.1	50.4	56.8	56.8	51.8	38.2	49.3	30.6	46.7	21.5
MFLVC (CVPR'22)	63.1	56.6	73.3	65.2	52.1	39.4	53.8	32.8	47.3	22.6
SEM (NeurIPS'23)	69.2	59.2	82.6	75.3	36.6	23.5	53.1	30.9	42.2	18.9
DIVIDE (AAAI'24)	60.9	53.8	64.3	57.9	46.5	27.0	52.3	33.5	45.6	23.0
SCMVC (TMM'24)	75.9	66.3	84.4	72.9	36.1	24.8	54.2	32.3	46.5	24.1
SSLNMVC (TMM'25)	64.4	58.3	82.1	72.8	44.8	32.3	51.2	33.0	46.4	24.0
DDMC	**76.7**	**68.8**	**90.3**	**82.7**	**60.9**	**40.9**	**58.7**	**36.5**	**49.5**	**28.3**
Ours vs BestCompared	**0.8↑**	**2.5↑**	**5.9↑**	**7.4↑**	**6.0↑**	**0.4↑**	**4.5↑**	**3.0↑**	**2.2↑**	**4.2↑**

Hyperparameters α and β of DDMC method on five datasets



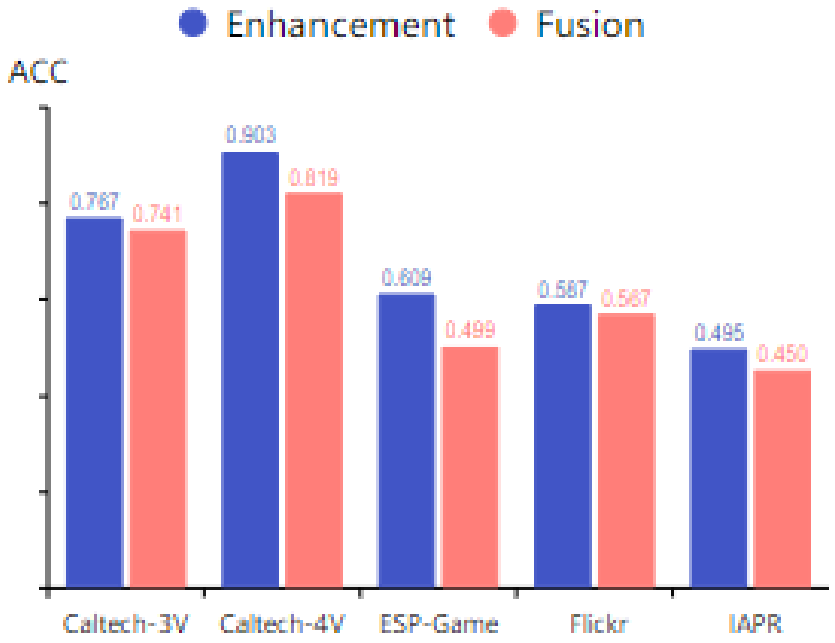
(a) Caltech-3V (b) Caltech-4V (c) ESP-Game (d) Flickr (e) IAPR

Ablation study of DDMC method on five datasets

Methods	Caltech-3V		Caltech-4V		ESP-Game		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
(1) $\mathcal{L}_{DDC}$	70.2	53.2	77.8	69.6	35.3	14.4	49.0	30.7	39.1	18.0
(2) $\mathcal{L}_{DDC} + \mathcal{L}_{FDL}$	71.4	61.3	79.4	73.5	54.9	35.5	55.2	35.4	48.8	28.1
(3) $\mathcal{L}_{DDC} + \mathcal{L}_{CDL}$	75.4	64.3	82.1	81.8	36.6	22.0	55.3	36.3	45.4	24.0
(4) All Modules (The Proposed Method)	**76.7**	**68.8**	**90.3**	**82.7**	**60.9**	**40.9**	**58.7**	**36.5**	**49.5**	**28.3**

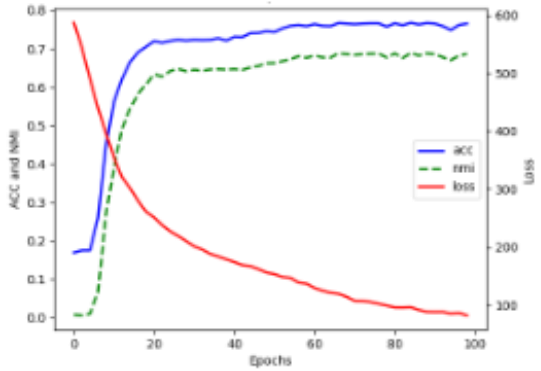
The experiments validate the significant contribution of each component in the proposed DDMC to the final clustering performance, fully proving its effectiveness.

Compared with enhancing the effectiveness of integration on five datasets

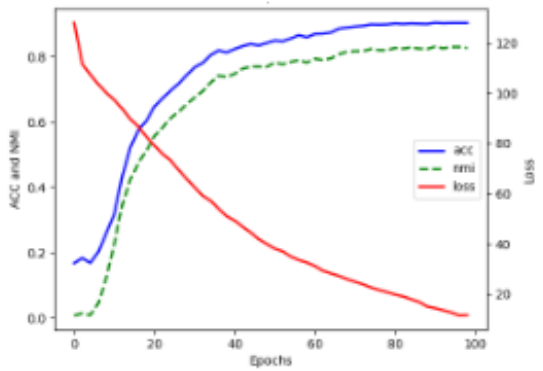


Experiments show that the enhancement strategy outperforms the fusion method on all datasets; especially on the ESP-Game dataset, the ACC is improved by about 11\%. This result shows that compared with modality fusion, dominant modality enhancement can significantly improve clustering performance, while the noise introduced in the fusion process will weaken the role of the dominant modality and reduce the clustering effect, further verifying the effectiveness of the DDMC.

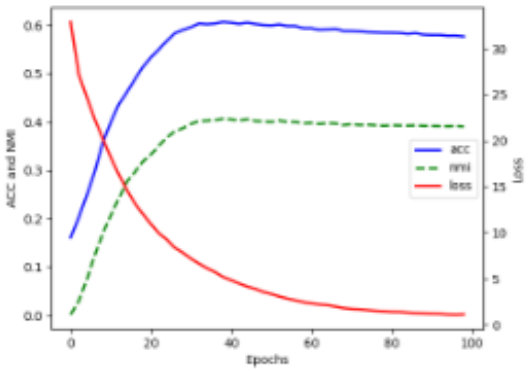
Convergence analysis of DDMC method on datasets



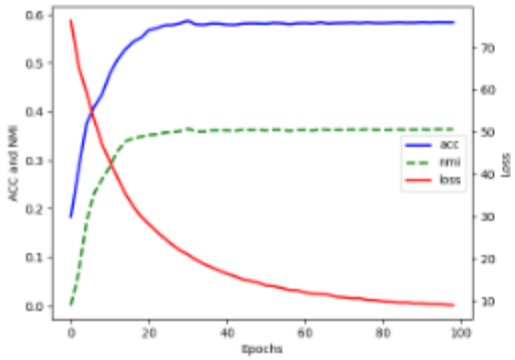
(a) Caltech - 3V



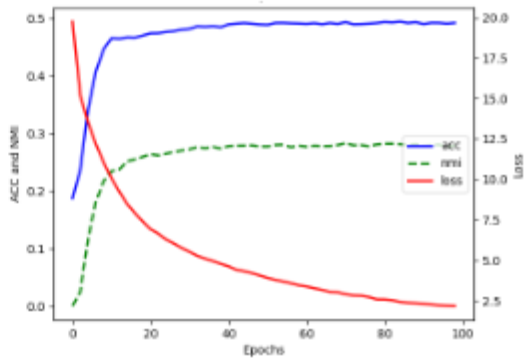
(b) Caltech - 4V



(c) ESP - Game

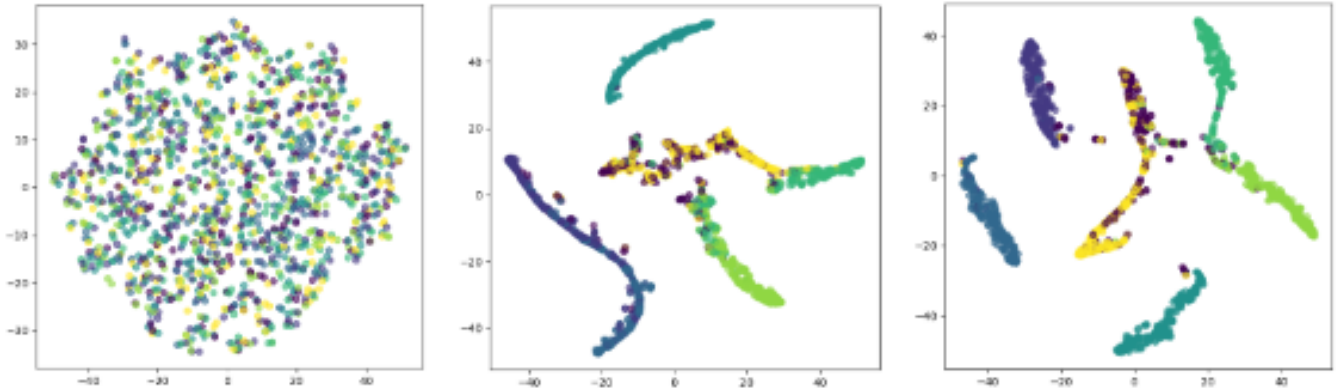


(d) Flickr

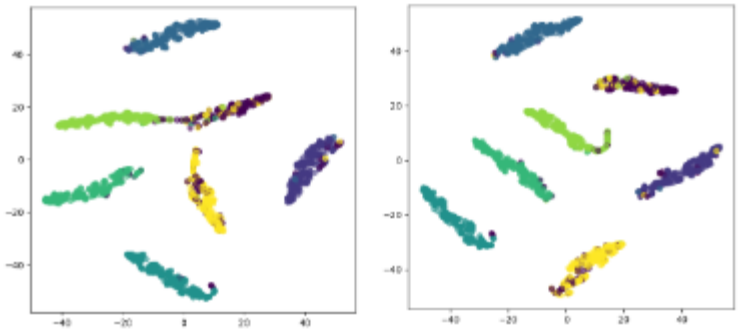


(e) IAPR

T-SNE visualization of Clustering results on five datasets



(a) 0 epoch (b) 25 epochs (c) 50 epochs



(d) 75 epochs (e) 100 epochs

Outline

- Problem background
- Previous works
- Our proposal
- Experiments
- **Conclusion**

Summary

This paper introduces a novel deep multi-modal clustering framework that leverages a dominant-modality enhancement strategy to mitigate noise from conventional feature-fusion. Rather than fusing all modalities indiscriminately, we identify the highest-quality modality as dominant, perform two-level diversity learning to extract diversity information from the remaining modalities, and augment the dominant modality accordingly. This strategy significantly improves the clustering performance, especially when the multi-modal data is unevenly distributed or has large quality differences.

Thank You!

Contact for communication:

ieshizhehu@zzu.edu.cn