Super Deep Contrastive Information Bottleneck for Multi-modal Clustering

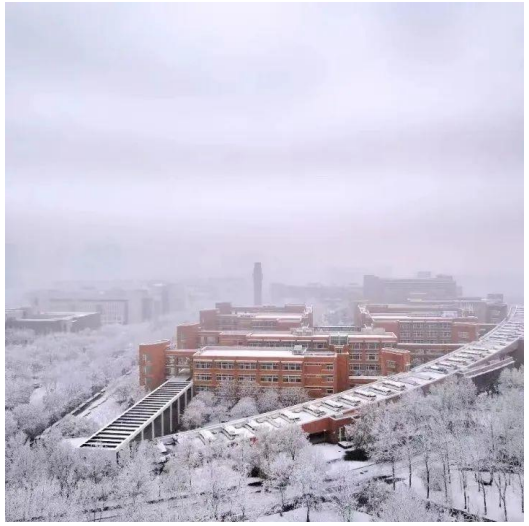
Zhengzheng Lou , Ke Zhang , Yucong Wu , Shizhe Hu†

School of Computer and Artificial Intelligence

Zhengzhou University

Zhengzhou, Henan, China

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



Tourist Spot

Outline

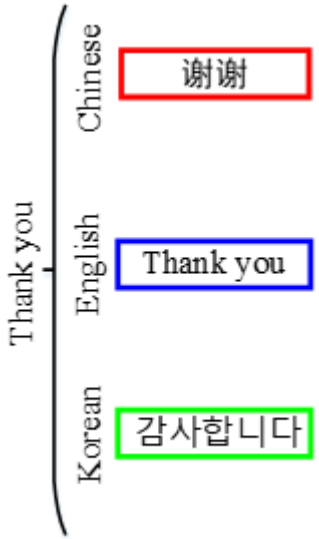
- Problem background
- Previous works
- Our proposal
- Experiments
- Conclusion

Outline

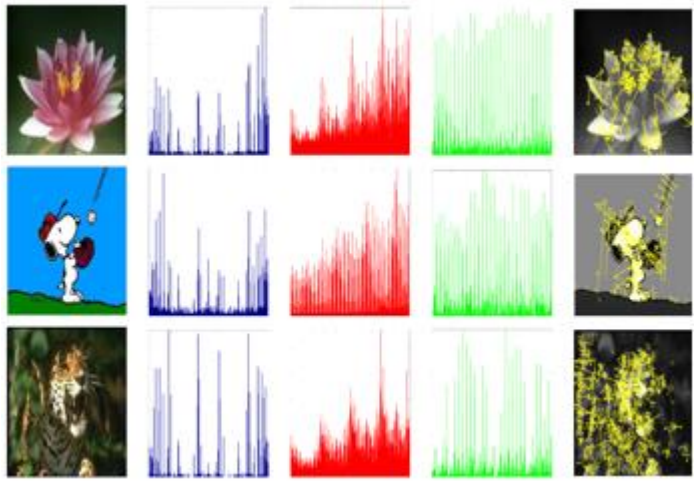
- **Problem background**
- Previous works
- Our proposal
- Experiments
- Conclusion

Characteristics of multi-modal datasets

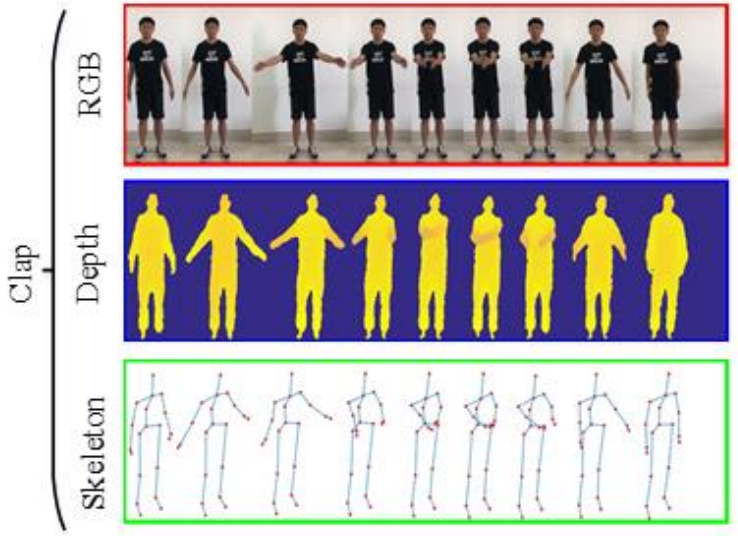
In Big Data era, many kinds of multi-modal data are emerging.



Multi-lingual Text



Multi-feature Image

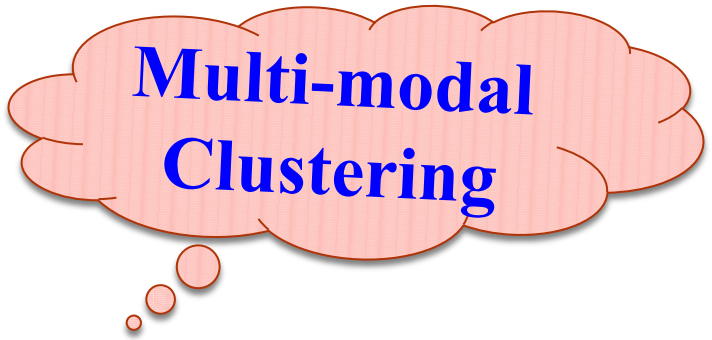


Multi-modal human action video

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

1. **Simply achieving consensus fails to capture complex latent information and interdependencies between modalities.;**
2. **Focusing only on single data information (e.g., clustering or features) ignores inherent latent structural information and modality heterogeneity.**

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-view clustering methods

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

  Wang et al. (Wang et al., 2021) combined the self-supervised t-SNE module with the self-expression layer to learn a shared low dimensional representation.

  Mao et al. (Mao et al., 2021) adopted the idea of contrastive learning to maximize the shared information between modalities and minimize the redundancy within each modality, achieving efficient clustering by integrating a multi-modal shared encoder with variational optimization.

  Rong et al. (Rong et al., 2022) utilized a variational autoencoder architecture based on the autoencoder and incorporated an attention mechanism to extract cluster-friendly representations from multi-omics data.

Limitations:

- *Fails to deeply understand the complex relationships within data samples across modalities, neglecting the close connections between the data.*

Outline

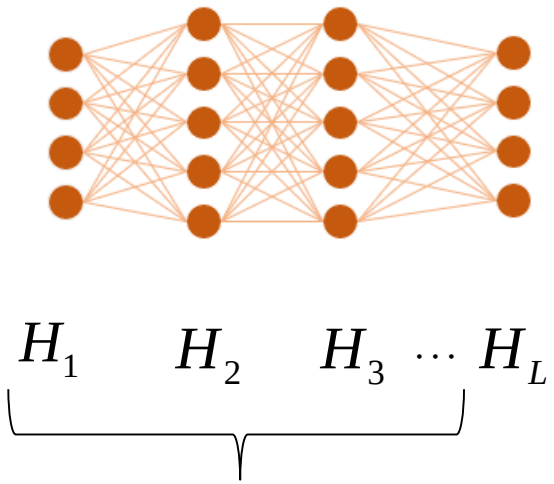
- Problem background
- Previous works
- **Our proposal**
- Experiments
- Conclusion

Our proposed method

- Super Deep Contrastive Information Bottleneck (SDCIB):
  - Hidden-layer Information Part;
  - Information Bottleneck Part;
  - Consistency Information Part.

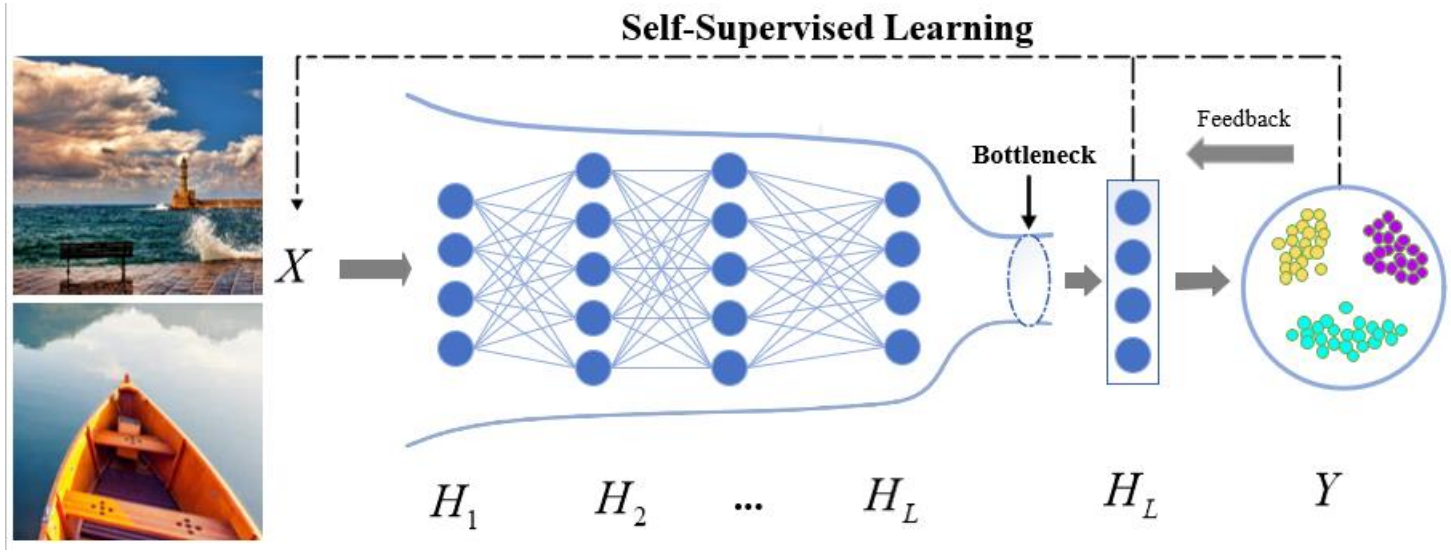
Hidden-layer Information Part

The features output by the hidden layer are rich in information and are explicitly used in the objective function.



Information Bottleneck Part

IB compresses the original features X^i to obtain a good compression representation H_L^i , so as to obtain a better clustering result Y_i .



$$\mathcal{L}_1 = \sum_{i=1}^{v} \sum_{l=1}^{L} I(X^i; H_l^i) - \beta I(Y^i; H_l^i)$$

Consistency Information Part

The consistency of feature level and cluster level is considered at the same time, and the hidden layer features are also included in the comparison scope.

$$\mathcal{L}_2 = \sum_{i=1}^{v}\sum_{j=1}^{v}\mathbb{I}_{i\neq j}[I(Y^i;Y^j) + \sum_{l=1}^{L} I(H_l^i;H_l^j)]$$

Objective function

We propose a novel superdeep contrastive information bottleneck method:

$$\begin{aligned}\mathcal{L}_{SDCIB} &= \alpha\mathcal{L}_{SIB} - (1-\alpha)\mathcal{L}_{Con} \\ &= \alpha\sum_{i=1}^{v}\sum_{l=1}^{L}[I(X^i;H_l^i) - \beta I(Y^i;H_l^i)] \\ &- (1-\alpha)\sum_{i=1}^{v}\sum_{j=1}^{v}\mathbb{I}_{i\neq j}[I(Y^i;Y^j) + \sum_{l=1}^{L} I(H_l^i;H_l^j)]\end{aligned}$$

α denotes the **balance parameter** between IB and consistency information

β denotes the **balance parameter** trading off the information compression and preservation

Advantages of the SDCIB

- The first to explicitly introduce the information contained in the encoder’s hidden layers into the loss function;
- Performs dual optimization by simultaneously considering consistency information from both the feature distribution and clustering assignment perspectives;

Optimization method

Algorithm 1 Algorithm for Optimizing the proposed SD-CIB

Input: Dataset with M modalities $\{X^i\}_{i=1}^m$, the number of clusters K, the parameter α, β.
Output: Final partition Y.
Random Initialization: Randomly initialize the parameters of M modality-specific encoders and $M * L$ mutual information estimators.
repeat
    Calculate L_{MINE} by Eq. (12)
    Optimize the parameters of mutual information estimators
    Obtain $I(Y^i; H_l^i)$ by mutual information estimators
    Calculate $I(X^i; H_l^i)$ by Eq. (8)
    Calculate $I(H_l^i; H_l^j)$ and $I(Y^i; Y^j)$ and by Eq. (11)
    Optimize the parameters of modality-specific encoders
until Convergence
Obtain partition Y^i
Obtain final partition Y by Eq. (2)

Outline

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

Datasets

Dataset	Modalities	Samples	Clusters	Dimensionality
Caltech-2V	2	1440	7	(40,254)
Event	3	1579	8	(1000,1000,1000)
IAPR	2	7855	6	(1200,500)
ESP	3	11032	7	(300,300,300)

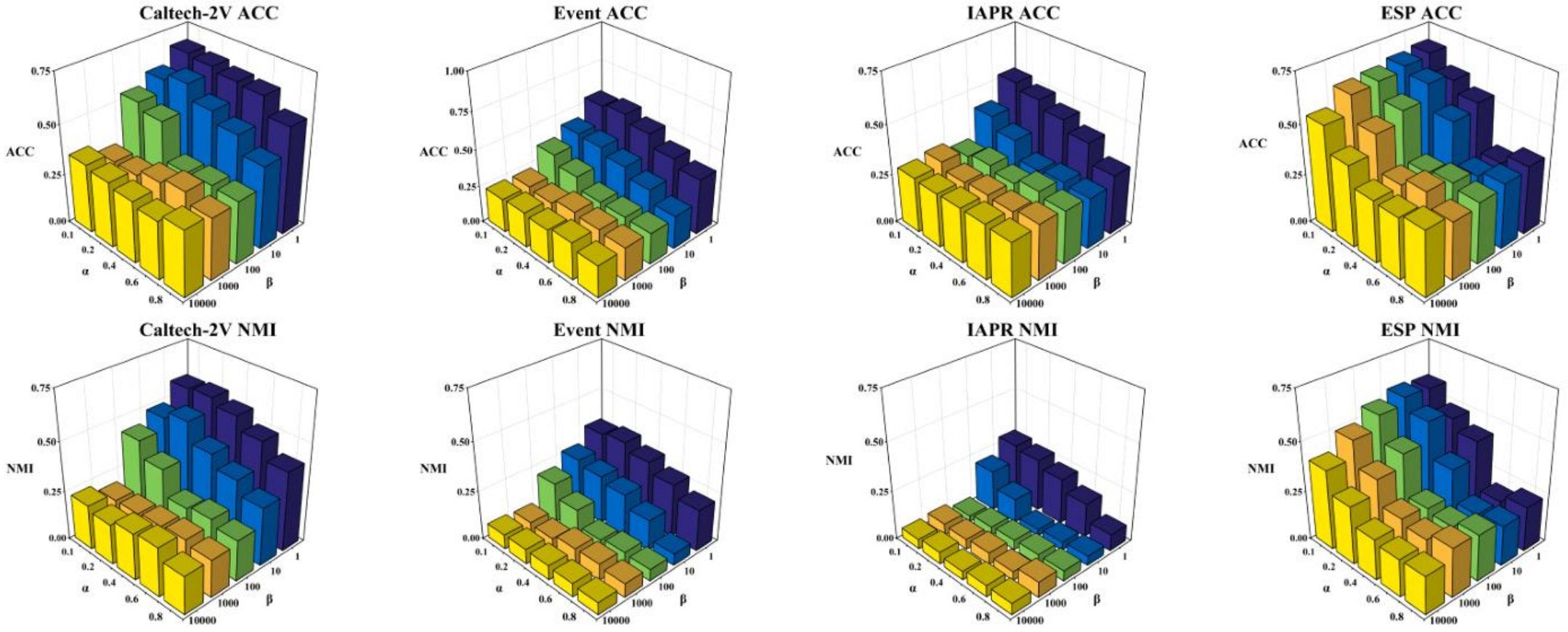
Compared methods

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) ONMSC: A multi-modal clustering method that integrates the neighborhood information of first-order and high-order laplacian matrices.
4) **Deep Clustering**:
   (1) 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.
   (2) MFLVC: A hierarchical feature learning clustering method that efficiently integrates multi-level feature learning and contrastive learning.
   (3) DealMVC: A clustering method that ensures the consistency of similar samples using a dual contrastive calibration network.
   (4) ICMVC: An end-to-end clustering method that handles missing data through multi-modal consistency transfer and graph convolutional networks, and combines contrastive learning.
   (5) DIVICE: 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.

Clustering results

Methods	Caltech-2V		Event		IAPR		ESP	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	41.6	30.5	34.7	20.7	38.9	17.2	48.4	33.5
Ncuts (TPAMI'00)	39.9	31.2	34.8	15.5	41.9	18.9	45.7	29.8
AmKM	46.4	31.4	28.7	11.6	40.4	17.0	35.0	20.7
AmNcuts (TPAMI'00)	42.8	5.2	35.2	20.3	42.2	18.9	32.5	19.0
CoregMVSC (NIPS'11)	49.2	39.6	35.5	22.2	35.1	18.4	45.2	30.7
RMKMC (IJCAI'13)	51.4	33.5	39.5	25.1	36.4	15.9	35.1	20.8
SwMC (IJCAI'17)	34.2	26.6	16.7	2.2	30.2	23.1	37.8	22.8
ONMSC (AAAI'20)	34.2	26.6	48.6	33.8	21.6	11.1	21.2	12.2
SiMVC (CVPR'21)	51.1	36.9	36.8	23.1	42.7	18.5	33.6	14.6
CoMVC (CVPR'21)	59.2	49.2	49.1	35.5	46.7	21.5	43.4	27.3
MFLVC (CVPR'22)	61.5	53.6	48.5	34.9	47.3	22.6	52.1	36.9
DealMVC (ACM MM'23)	47.6	37.9	26.5	9.3	35.0	10.8	43.4	27.4
ICMVC (AAAI'24)	49.6	37.9	36.4	30.3	37.1	16.8	46.7	30.0
DIVICE (AAAI'24)	64.1	52.9	31.4	12.4	45.6	23.0	47.2	28.8
SDCIB	**67.5**	**59.2**	**56.5**	**36.4**	**52.9**	**28.7**	**61.4**	**44.7**
Ours vs Best Compared	3.4↑	5.6↑	7.4↑	0.9↑	5.6↑	5.6↑	9.3↑	7.8↑

Hyperparameters α and β of SDCIB method on four datasets



Number of encoder layers of SDCIB method on four datasets

Hidden layers	Caltech-2V		Event		IAPR		ESP	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
2	62.6	54.8	50.1	30.3	52.7	27.2	**62.1**	42.5
3	64.4	54.2	55.0	**36.6**	51.8	28.0	61.6	**44.7**
4	**67.5**	**59.2**	**56.5**	36.4	**52.9**	28.7	61.4	**44.7**
5	65.7	58.3	47.1	30.3	49.4	26.1	58.8	42.1
6	64.5	57.1	50.5	31.3	52.4	**38.5**	62.0	43.8
7	65.3	58.3	44.5	27.6	47.6	27.3	61.2	43.7

The results indicate that having more or fewer layers does not necessarily lead to better performance.

Ablation study of SDCIB method on four datasets

Methods	Caltech-2V		Event		IAPR		ESP	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
removing $I(Y^i;Y^j)$	41.8	33.2	30.3	15.5	38.5	20.1	36.3	27.6
removing $I(H_l^i;H_l^j)$	63.1	52.2	45.0	25.3	48.8	23.5	**62.3**	43.4
removing $I(Y^i;Y^j)$ and $I(H_l^i;H_l^j)$	32.2	23.6	21.6	9.1	30.0	6.8	27.1	16.3
SDCIB	**67.5**	**59.2**	**56.5**	**36.4**	**52.9**	**28.7**	61.4	**44.7**

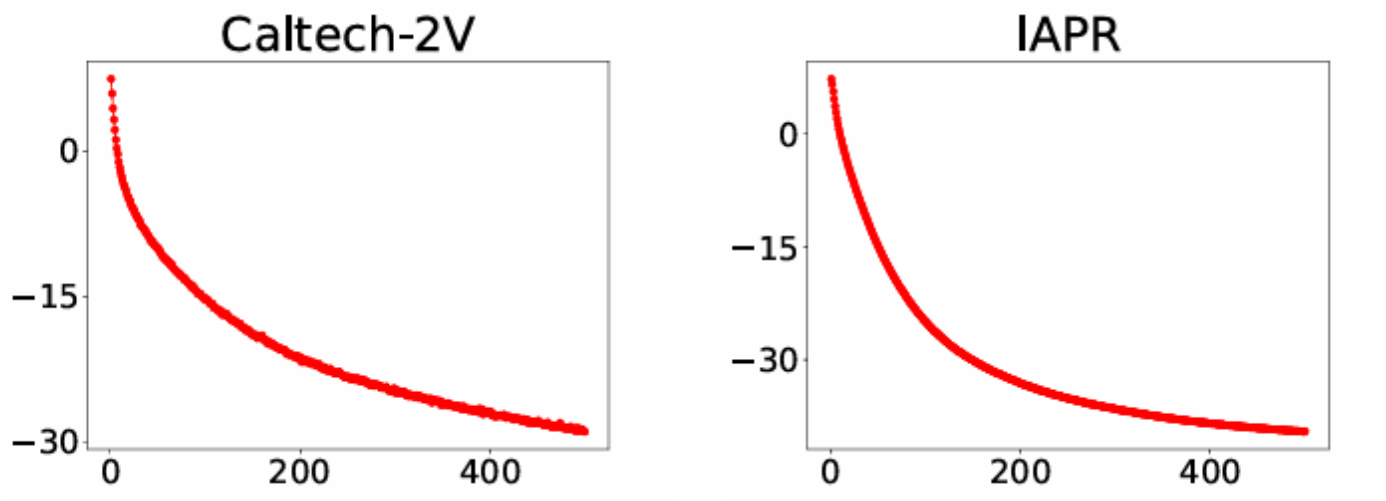
The experiments validate the significant contribution of each component in the proposed SDCIB to the final clustering performance, fully proving its effectiveness.

Necessity of hidden layer information of SDCIB method on four datasets

Methods	Caltech-2V		Event		IAPR		ESP	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
SDCIB-no-hidden	63.6	55.8	45.9	29.1	48.9	26.9	57.9	40.0
SDCIB	**67.5**	**59.2**	**56.5**	**36.4**	**52.9**	**28.7**	**61.4**	**44.7**
Improvement	3.9↑	3.4↑	10.6↑	7.3↑	4.0↑	1.8↑	3.5↑	4.7↑

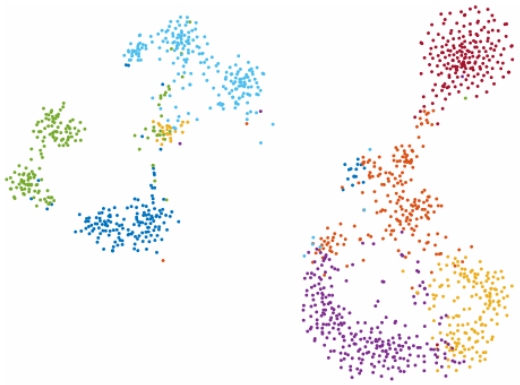
This experiment indicate that fully mining and utilizing hidden layer information helps to more deeply explore the intrinsic relationships and latent structures among modalities, thereby enhancing the accuracy and robustness of clustering results.

Convergence analysis of SDCIB method on datasets



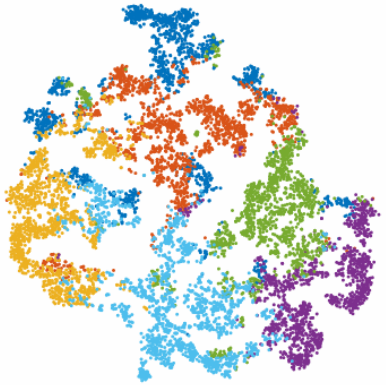
The proposed SDCIB demonstrates rapid convergence within a certain range, which not only validates the effectiveness of the proposed SDCIB but also demonstrates the reliability and stability of the proposed SDCIB.

T-SNE visualization of Clustering results on four datasets

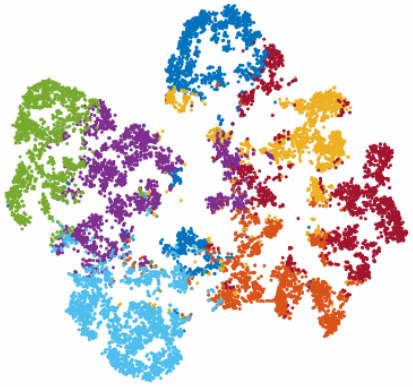


(a) Caltech-2V

(b) Event



(c) IAPR



(d) ESP

Outline

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

Summary

- Create a novel super deep contrastive information bottleneck (SDCIB) for multi-modal clustering;
- The proposed SDCIB not only incorporates the rich information contained in the encoder's hidden layers into the clustering process, but also performs dual optimization from two consistency information perspectives: feature distribution and clustering assignment;
- Our approach achieves state-of-the-art performance.

Thank You!

Contact for communication:

ieshizhehu@gmail.com