Self-supervised Trusted Contrastive Multi-view Clustering with Uncertainty Refined

Shizhe Hu, Binyan Tian, Weibo Liu, Yangdong Ye *

School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
ieshizhehu@gmail.com, binyan@gs.zzu.edu.cn, liuweibo@stu.zzu.edu.cnl.com, ieydye@zzu.edu.cn

Abstract

Multi-view clustering (MVC), especially contrastive MVC, has demonstrated promising potential in many fields and practical scenarios. However, existing contrastive MVC methods still ignore the reliability of clustering results and the impact of false negative pairs, which limits the application of methods in critical security areas. To solve the above challenges, we propose a Self-supervised Trusted Contrastive Multi-view Clustering with Uncertainty Refined (STCMC-UR) method, which integrates clustering results and uncertainty learning to guide the self-supervised contrastive learning. First, the evidence of a specific view is generated in the evidence generation module. Afterwards, the belief masses and uncertainty of each view are learned and we fuse multiple views with the Dempster-Shafer theory to generate the final clustering result and the uncertainty of the view. Different from existing methods, with the clustering result and uncertainty generated by the fusion, we design a feature-level uncertainty-refined self-supervised contrastive learning module, where the pseudo-label is selectively employed in each iteration to conduct more accurate contrastive learning. As a result, the modules are mutually beneficial, which is conducive to more effective feature learning and clustering structure discovery, and more accurate learning results are obtained. Extensive experiments on five datasets show that the proposed method has significant improvements in effectiveness compared with the latest methods.

Code — https://github.com/ShizheHu.

Introduction

In recent years, multi-view clustering (MVC) has achieved satisfactory clustering performance (Li et al. 2019; Tao et al. 2019) by exploring and utilizing the consistency and complementary information between multiple views, and has been widely adopted in various practical scenarios (Hu et al. 2022; Han et al. 2022a).

Motivation. With the rapid development of deep learning (Mao et al. 2021; Han et al. 2024), deep neural networks based MVC methods have gained attention due to their superior capabilities in feature extraction and clustering performance compared to traditional methods. Among these MVCs, contrastive learning based MVC demonstrates excellent performance by learning deep feature representations in a self-supervised manner (Hu et al. 2023; Chen et al. 2024; Wang et al. 2020). For example, in the work (Lu et al. 2024a), it utilized similarity information within and across the modalities to learn consistent representations. Recent method in (Hu et al. 2024) further optimized feature representation and clustering results by incorporating contrastive learning with dual correlation learning. However, the current deep contrastive MVCs still face several challenges. (1) First, most of them select the same samples between views as positive pairs while treating other samples as negative pairs. This often overlooks the numerous false negatives within the negative sample pairs that need to be eliminated, leading to inaccurate contrastive learning (Zhao et al. 2019). (2) Existing methods typically ignore the varying importance of multiple views or simply attain unchanged view weights, leading to limited generalization and learning ability. (3) Although these methods have been proven effective in experiments, they often lack reliability and robustness (Gawlikowski et al. 2023; Wu et al. 2022). Most rely on deterministic factors to define the output results, but they ignore the estimation of uncertainty, which makes it impossible to guarantee the reliability of the decision, thus limiting their practical application.

Recently, the characterization of model uncertainty has received extensive attention (Zheng et al. 2023; Zhu et al. 2024), especially in safety-critical fields such as medical diagnosis and self-driving, where understanding the reliability and confidence of algorithmic decisions is crucial (Pedersen et al. 2017). A series of uncertainty-based learning has been proposed, including Bayesian neural networks (BNN) (Blundell et al. 2015), confidence learning (Han et al. 2022b), and evidence-based learning (Audun 2001). For example, BNN estimates uncertainty by learning the distribution of weights. However, due to the surge in the size of neural network parameters, BNN is computationally expensive. MC-Dropout (Gal and Ghahramani 2016) alleviates this problem to a certain extent by dropout sampling. Confidence learning is committed to capturing and optimizing the confidence of the results through a carefully designed calibration process to ensure its reliability. Different from the Bayesian theory, Dempster-Shafer theory is a classic evidence theory first proposed by Dempster (Dempster 2008)

*Corresponding author

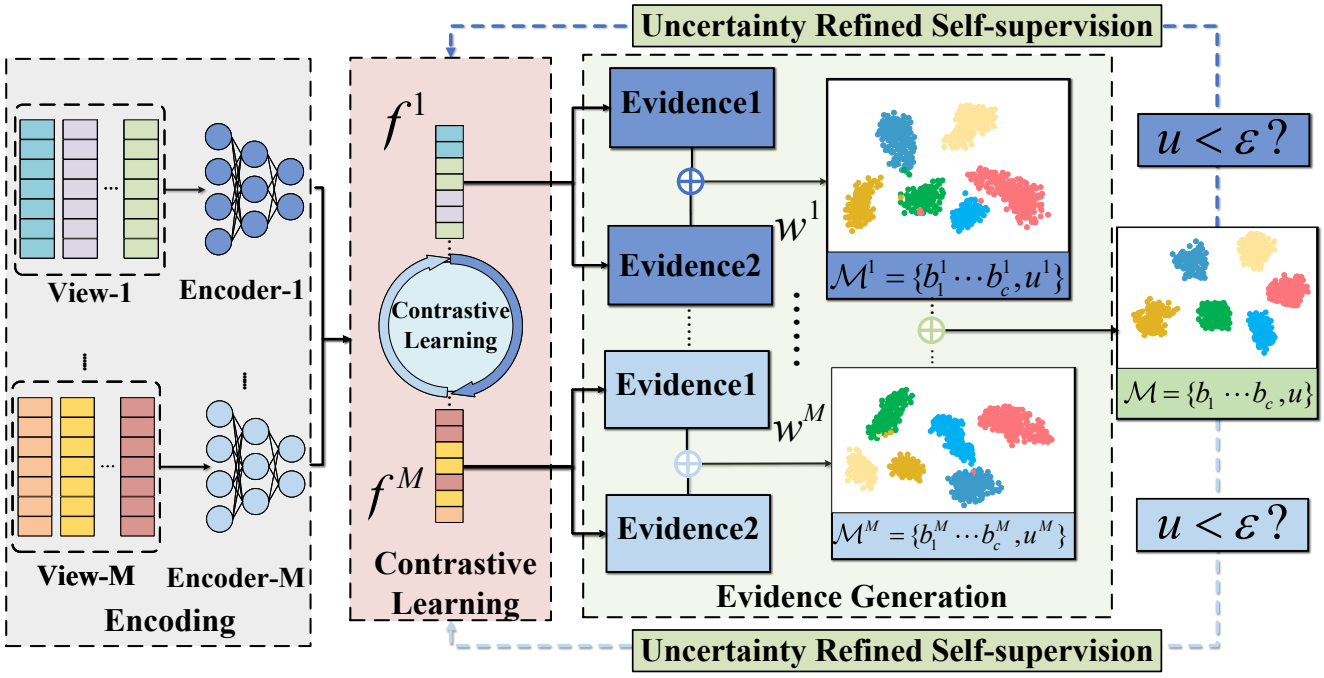

Figure 1: The proposed STCMC-UR. It consists of three modules: (1) The feature representation f^v of the v-th view is obtained through the encoder, and the evidence representation e^v is obtained by fusing two BBAs. (2) The final pseudo-label and uncertainty are obtained by fusing multi-views and the view weight w^v is quantified by clustering results to adjust e^v. (3) By comparing the uncertainty u and the threshold ε, the highly confident pseudo labels are automatically selected to self-supervise the feature-level contrastive learning.

and Shafer, and is extended to multi-view setting by (Han et al. 2022c). This theory fuses the shared parts between views, while ignoring conflicting beliefs through normalization factors. However, most research relies on only one evidence generation method, which leads to biased allocation output (Shao, Dou, and Pan 2024). Moreover, most of the trusted methods depend on manually annotated datasets, limiting their applicability to unlabeled datasets encountered in real-world scenarios. In addition, there is currently no effective solution for how to leverage the uncertainty obtained after fusion to guide the model in more reliable training.

Contribution. To overcome the aforementioned challenges, we propose a novel MVC framework, named STCMC-UR, which integrates Dempster-Shafer theory to guide contrastive learning under unsupervised conditions, as shown in Figure 1. Specifically, reliance on a single Basic Belief Assignment (BBA) approach may lead to biased allocation results, we here use two BBAs to achieve better generalization performance. The outcome of the neural network is linked to the evidence, and then the two BBAs are fused using the Dirichlet distribution combined with Dempster-Shafer theory to obtain the evidence for each view. After that, the Dirichlet is employed in combination with the Dempster-Shafer theory to fuse at the multi-view level to reach the final clustering result and uncertainty. Considering the different weights of each view, we quantify the view weights with the clustering results obtained to adjust the evidence, so as to make full utilization of the high-quality parts of the input data. In addition, unlike the existing deep MVCs, we flexibly integrate the uncertainty and pseudo-label to guide feature-level contrastive learning where a threshold is set for uncertainty to filter high-quality pseudo labels. For samples with a value greater than the threshold, the confidence of the clustering result is considered too low, so the pseudo-label is not used to guide contrastive learning. Therefore, the pseudo-label is selectively integrated to remove inaccurate negative instance pairs, making the feature-level contrastive learning more accurate. We apply STCMC-UR to different datasets and the results demonstrate that our method achieves outstanding performance compared to the latest methods. The major contributions are summarized as:

- A novel trusted MVC framework is proposed, which incorporates the uncertainty into contrastive MVC and reaches more reliable contrastive learning and clustering result. To our knowledge, this is the first work that considers both uncertainty-refined contrastive learning and trusted learning.
- The view evidence is obtained by fusing two BBAs, which effectively avoids the problem of over-confidence, making the model more generalizable.
- An uncertainty-refined self-supervised contrastive learning mechanism is designed. The uncertainty is combined with pseudo-label to guide the feature-level contrastive learning, making the contrastive learning more accurate

and thus leading to more discriminative representation.

The Proposed Method

Problem Formulation

Assume that $\{X^v\}_{v=1}^M$ represents the feature matrix of M views extracted based on initial data, where the sample of the v-th view is represented as $X^v = \{x_n^v\}_{n=1}^N \in \mathbb{R}^{N\times d^v}$, and its feature is represented as f^v, where N represents the number of instances in each view, and d^v represents the dimension of the v-th view. Our aim is to cluster the data across multi-views, where c denotes the number of clusters.

Overview of Objective Function

In general, STCMC-UR jointly optimizes three parts of loss: (1) the loss of the evidence generation stage; (2) the loss of view fusion; and (3) the loss of feature-level contrastive learning. The three modules are jointly executed, making the modules mutually beneficial. Accordingly, the total objective function is given as follows:

$$\mathcal{L} = \mathcal{L}_C + \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{CL}. \tag{1}$$

where $\alpha, \beta\in(0, 1)$ are the balancing parameters of the loss. The details of each loss are shown in the following sections. A brief introduction is shown as:

$\mathcal{L}_C$: A loss function for evidence generation is designed, which represents the loss of obtaining beliefs from each view during training.

$\mathcal{L}_{CE}$: In order to minimize the conflict between different viewpoints in the multi-view fusion process, the shared information and cluster correlation are discovered as much as possible to ensure the consistency of the results.

$\mathcal{L}_{CL}$: This part represents the alignment of features between different views via contrastive learning. In each iteration, by comparing the uncertainty and the threshold, the pseudo-label with higher credibility are selected in a self-supervised manner to guide the contrastive learning, thereby enabling the model to learn higher quality representations.

Evidence Generation Module

In this section, we detail how to integrate evidence theory to form opinions on specific views. In order to generate beliefs, the popular practice is to utilize the Softplus function to connect the outcome of the neural network to quantify the view evidence. Unlike softmax, which often outputs over-confident results (Sensoy, Kaplan, and Kandemir 2018), the results produced by the Softplus function are considered more reliable. However, each view of the sample contributes to different degrees, and excessive reliance on a single BBA method may result in diminished generalization performance for the model. Therefore, inspired by (Shao, Dou, and Pan 2024), this paper fuses the output of the neural network through two BBAs to generate more robust view evidence. The BBA methods are defined as follows:

Definition 1 (BBA-SP) *BBA-SP adopts the conventional evidence generation method and utilizes the Softplus function to obtain evidence:*

$$e_k^1 = Softplus(o) = ln(1 + e^o). \tag{2}$$

Definition 2 (BBA-SM) *BBA-SM derives the evidence by the reallocation of the outcome of deep neural network with the softmax function:*

$$e_k^2 = E \cdot softmax(o) = \frac{E \cdot e^{o_k}}{\sum_{j=1}^c e^{o_j}}, \tag{3}$$

where $E = \sum_j^c o_j$ represents the total amount of evidence.

By introducing subjective logic, belief mass is linked to Dirichlet distribution parameters. The Dirichlet distribution parameter α is defined as: $\alpha = e + 1 = \sigma(o) + 1$, where σ represents a non-negative activation function. Thus the Dirichlet distribution parameter satisfies $\alpha_k \geq 1$ when $c \geq k \geq 1$. Dirichlet combine evidence learning from different angles and Dempster-Shafer theory to learn the class probability and uncertainty of the model, making the model reliable and robust. The belief mass and uncertainty are obtained by the following formula:

$$b = \frac{e}{S} = \frac{(\alpha - 1)}{S}, u = \frac{c}{S}, \tag{4}$$

where $S = \Sigma_{j=1}^K \alpha_j$ indicates the Dirichlet intensity related to the Dirichlet distribution and satisfies

$$u + \sum_{k=1}^{c} b_k = 1, u \geq 0, b_k \geq 0. \tag{5}$$

which indicates a whole distribution of beliefs and uncertainty. BBA-SP preserves the initial evidence from the neural network and reach more reliable predictions. However, this approach is problematic when the inputs with a high level of confidence are given (Hein, Andriushchenko, and Bitterwolf 2019). BBA-SM method generates view evidence through the softmax function, which can lead to over-confidence but meanwhile offers greater certainty for some predictions. Through Dempster-Shafer theory, the complementary information of the two methods can be fully utilized to fuse multi-source evidence. The two pieces of evidences of the v-th view of the n-th instance, we denote them by $e_{n,1}^v$ and $e_{n,2}^v$, belief masses by $b_{n,1}^v$ and $b_{n,2}^v$, and uncertainty by $u_{n,1}^v$ and $u_{n,2}^v$, and the two BBAs are fused by the following Dempster-Shafer theory equation:

$$(b_n^v)^{\oplus} = b_{n,1}^v \oplus b_{n,2}^v = \frac{1}{\kappa}(b_{n,1}^v \odot b_{n,2}^v + b_{n,1}^v u_{n,2}^v + b_{n,2}^v u_{n,1}^v), \tag{6}$$

$$(u_n^v)^{\oplus} = \frac{1}{\kappa} u_{n,1}^v u_{n,2}^v, \tag{7}$$

$$(S_n^v)^{\oplus} = \frac{c}{(u_n^v)^{\oplus}}, \tag{8}$$

where $\oplus$ represents the Dempster-Shafer theory, $\odot$ represents the Hadamard (elementwise) product, and the normalization coefficient is denoted as κ based on the Eq. (5). Thus, we can obtain the view evidence after two BBAs:

$$(e_n^v)^{\oplus} = (S_n^v)^{\oplus} \cdot (b_n^v)^{\oplus}. \tag{9}$$

We integrate the fused e_n^v as the evidence of the v-th view of the n-th instance for subsequent task, so that the obtained belief mass and uncertainty can offer more accurate and robust estimation. In this part, we give an effective evidence generation loss function, which strengthens

compact feature representation while ensuring the separability between different clusters. The loss consists of three parts, the first of which derives from the generalization of the Cauchy-Schwartz (CS) divergence across multiple densities. To guarantee the improvement of the sampling method, we use a data-driven method (Kampffmeyer et al. 2019) to optimize it as follows:

$$\mathcal{L}_1 = \frac{1}{c}\sum_{i=1}^{c-1}\sum_{j>i}\frac{\delta_i^T\mathbf{K}\delta_j}{\sqrt{\delta_i^T\mathbf{K}\delta_i\delta_j^T\mathbf{K}\delta_j}}, \tag{10}$$

where $\mathbf{K}$ represents the Gaussian kernel matrix and the column of clustering result is represented by the vector δ_i.

The second part applies geometric structure to CS divergence by forcing cluster partitions to simplex corners to avoid trivial solutions, as follows:

$$\mathcal{L}_2 = \frac{1}{c}\sum_{i=1}^{c-1}\sum_{j>i}\frac{\lambda_i^T\mathbf{K}\lambda_j}{\sqrt{\lambda_i^T\mathbf{K}\lambda_i\lambda_j^T\mathbf{K}\lambda_j}}, \tag{11}$$

where λ_i indicates the i-th column of the matrix $B = [B_{ab}] = \mathrm{e}^{(-||\alpha_a - e_b||^2)}$, the b-th corner of the simplex is represented by e_b.

The third term forces the output results to be orthogonal, and the formula is:

$$\mathcal{L}_3 = triu(G^TG), \tag{12}$$

where $triu(G^TG)$ represents the total of the elements located in the upper triangular portion of the G^TG matrix.

Thus, the loss is given by the following formula:

$$\mathcal{L}_C = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{13}$$

Multi-view Fusion

In this part, we focus on the aggregation across views. By using Dirichlet distribution to model the class probability distribution and then integrating it with Dempster-Shafer theory, different view opinions are used to generate belief mass and uncertainty of multi-view fusion at the evidence level.

First, using the fused evidence, the view-specific opinion of the i-th sample can be obtained by Eq. (4). Dempster-Shafer theory provides a method to combine joint evidence from different sources. Using this rule, multiple opinions are aggregated to produce a joint opinion. For the convenience of explanation, take two view opinions $\mathcal{M}^1 = \{\{b_k^1\}_{k=1}^c,\ u^1\}$ and $\mathcal{M}^2 = \{\{b_k^2\}_{k=1}^c,\ u^2\}$ as an example, and the joint opinion is calculated as follows:

$$\mathcal{M} = \mathcal{M}^1 \oplus \mathcal{M}^2, \tag{14}$$

where the belief mass and uncertainty calculation of the combination are shown in the above Eq. (6)-(8).

Through the Dempster-Shafer theory, it is ensured that when the opinions of the two views are both large or both small, the uncertainty of the fusion opinion is either large or small; when the uncertainty of a certain view is very low, the final result will depend on the view with high confidence. Therefore, the joint opinion reduces the risk opinions while utilizing valuable opinions, so it is credible and interpretable. Therefore, for the case of multi-view, then we have

$$\mathcal{M} = \mathcal{M}^1 \oplus \mathcal{M}^2 \oplus \ldots \mathcal{M}^M. \tag{15}$$

Finally, after view fusion, the final clustering result is reached by the maximum class probability. At the same time, in order to minimize the degree of conflict between opinions during the fusion process and ensure that the results of different opinions tend to be consistent, inspired by (Xu et al. 2024), we introduce the consistency loss calculation of the instance $\{x_n^m\}_{m=1}^M$ as follows:

$$\mathcal{L}_{CE} = \frac{1}{M-1}\sum_{a=1}^{M}\left(\sum_{b\neq a}^{M}\Phi(\eta_n^a, \eta_n^b)\right), \tag{16}$$

where $\Phi(\eta_n^a, \eta_n^b) = \Phi_p(\eta_n^a, \eta_n^b)\cdot\Phi_c(\eta_n^a, \eta_n^b)$ represents the degree of conflict between two views η_n^a and η_n^b of n-th instance. Specifically, η is an ordered triple consisting of view point $\mathcal{M}$ and $\boldsymbol{\rho}$, where $\boldsymbol{\rho} = (\rho_1, ..., \rho_c)^\top$ reflects the prior probability distribution on each clusters and the initial value is 1/c indicating that each cluster contains a comparable amount of data instances. $\Phi_p(\eta_n^a, \eta_n^b)$ and $\Phi_c(\eta_n^a, \eta_n^b)$ represent the projection distance and conjunction certainty between η_n^a and η_n^b respectively, which is calculated by $\Phi_p(\eta_n^a, \eta_n^b) = \{\sum_{i=1}^{c}\left|p_i^a - p_i^b\right|\}/2$ where $p_i = \alpha_i/S$ and $\Phi_c(\eta_n^a, \eta_n^b) = (1-u^a)(1-u^b)$.

In addition, considering the different view contributions, we utilize the final clustering results to learn the weights of the views and further adjust the evidence so that the evidence representation can fully exploit the valuable information of the view while reducing the contribution of risky information. Specifically, we first assign an initial value of $1/M$ to the view weight, and then update the view weight by optimizing the objective function, just like updating the network parameters. Then, the softmax function is utilized to obtain the view weight for each period. This is simple and effective, and it will not make the entire network very complicated. The weight of the i-th view is represented by w^i, then we have

$$\sum_{i=1}^{m} w^i = 1(w^i > 0). \tag{17}$$

Uncertainty-refined Self-supervised Contrastive Learning

To further discover the consistency and complicated relationships among views, a feature-level contrastive learning module is designed to learn the common semantics between views by aligning the feature distribution. Contrastive learning draws the same instance across views as positive pairs, while treating other instances as negative pairs. It aims to maximize the similarity between positive ones, while minimizing it among negative ones to remove the feature noise of the samples and retain the consistent feature information between the same samples. We adopt the usual practice to quantify the pairwise similarity between two instances via the cosine distance:

$$s_{ij}^{uv}\left(f_i^u, f_j^v\right) = \frac{\left(f_i^u\right)^T f_j^v}{\left\|f_i^u\right\|\cdot\left\|f_j^v\right\|}, \tag{18}$$

where f_i^u and f_j^v represent the i-th and j-th sample representations from the u-th and v-th views respectively.

Simply treating samples within the same cluster as negative pairs may result in inaccurate contrastive learning (Lin et al. 2022; Wang et al. 2022; Hu et al. 2024). In fact, there are some false negative sample pairs in negative sample pairs, which may alienate the original positive samples in model training, and fail to fully and correctly learn similar representations in the same cluster, thus affecting the clustering performance. To this end, we try to integrate the pseudo-label and uncertainty obtained by view fusion, and guide feature-level contrastive learning by selectively using pseudo-label. The maximum clustering assignment probability obtained by multi-view fusion is regarded as a pseudo-label, combined with the generated uncertainty u. We define a fixed threshold ε uniformly set on all data sets. When $u \geq \varepsilon$, we believe that the clustering result of the sample has low confidence, so its pseudo-label is not used. In the remaining high-confidence samples, the clustering results are utilized to remove false negative pairs in CL, thereby selecting higher-quality pseudo-label to better learn distinguishable features and improve clustering performance.

Thus, in order to further optimize the similarity between paired samples, the NT-Xent contrastive learning loss is extended to the multi-view setting for joint optimization:

$$\mathcal{L}_{CL} = \frac{1}{NM(M-1)} \sum_{i=1}^{N} \sum_{u=1}^{M} \sum_{v=1}^{M} \mathbb{1}_{u \neq v} l_i^{uv}, \tag{19}$$

where $\mathbb{1}_{u \neq v} = 1$ if $u \neq v$, and l_i^{uv} is calculated by

$$l_i^{uv} = -\log \frac{e^{s_{ii}^{uv}(f_i^u, f_i^v)/\tau_1}}{\sum_{j=1}^{n} \mathbb{1}_{[pred_i \neq pred_j]} \cdot e^{s(f_i^u, f_j^v)/\tau_1}}, \tag{20}$$

where τ_1 is a temperature hyper-parameter, and pseudo-label for data points i and j are denoted by $pred_i$ and $pred_j$, respectively.

Optimization

The modules in our method are jointly optimized in an end-to-end fashion, and the losses of each part are balanced by α and β, and finally a promising clustering result is achieved. The details are shown in Algorithm 1.

Differences with Related Methods

In this section, we will further elaborate on the differences between the proposed method and existing contrastive learning methods. First, this is the first study to integrate uncertainty learning into contrastive multi-view clustering, resulting in more discriminative features and improved clustering accuracy. Second, unlike most methods, we use pseudo-labels to remove false negative pairs in each iteration, improving feature alignment and learning accuracy. Thus, STCMC-UR enhances contrastive learning by addressing clustering reliability and false negatives.

Complexity Analysis

Assume the dimensionality of the original multi-view samples the same, denoted as D, and the dimensionality of each layer in the multilayer perceptron network the same, denoted as P. Additionally, the number of layers in the multilayer perceptron network is denoted as L, the maximum number of neurons in the hidden layer of the encoder as Z, the dimensionality of the fusion features as E. We give the time complexity of optimizing each module as follows. Specifically, time cost of optimizing the L_{CL} module is $\mathcal{T}_{CL} = O(MNDP + MN(L-1)P^2 + M^2)) \approx O(NMDPL + M^2)$. The time complexity for L_{CE} module is $\mathcal{T}_{CE} = O(MN(c+1) + M^2N)$. Finally, the time complexity for L_C module is $\mathcal{T}_C = O(Mc^2)$. Thus, the overall time complexity is $\mathcal{T}_{total} = K(\mathcal{T}_{CL} + \mathcal{T}_{CE} + \mathcal{T}_C)$, where K is the number of training iterations.

Algorithm 1: Algorithm for STCMC-UR

Input: Multi-view data with M views, number of clusters c, hyper-parameters α, β and the threshold ε.
Output: The cluster results.
Initialize the parameters of the deep neural network.
Calculate the evidence for each view by Eq. (9).
Compute $\mathcal{L}_C$ by Eq. (13).
Multi-view fusion by Eq. (15).
Compute the view weight automatically.
Compute $\mathcal{L}_{CE}$ by Eq. (16).
Compute the $\mathcal{L}_{CL}$ by Eq. (19).
Jointly optimize the overall loss by Eq. (1).
return obtaining the final clustering result.

Dataset	# View	# Samples	# Clusters	# Dimensionality
Caltech-2V	2	1400	7	40/254
Caltech-3V	3	1400	7	45/254/928
COIL20	3	1440	20	$1 \times 128 \times 128$
Event8	3	1579	8	1000/1000/1000
NUS22	2	10155	22	1000/1000

Table 1: Details of various kinds of multi-view datasets.

Experiment

Experimental Settings

Datasets. Five frequently-used datasets are selected in the experiments shown in Table 1, and a detailed introduction to them are: Caltech[1] image dataset has two multi-view versions: **Caltech-2V** is categorized into 7 classes with a total of 1400 images. It contains two kinds of features. **Caltech-3V** includes the same classes and images compared to Caltech-2V, but introduces an additional feature. **COIL20**[2] dataset contains grayscale images categorized into 20 clusters. Each object is photographed at 5-degree intervals, resulting in 72 images per object. **Event8**[3] contains 8 sports event categories with 1,579 samples and three features. It poses challenges due to significant background variability

[1]https://data.caltech.edu/records/mzrjq-6wc02

[2]http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

[3]http://vision.stanford.edu/lijiali/event_dataset/

Method	Caltech-2V		Caltech-3V		COIL20		Event8		NUS22	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	41.6	30.5	46.3	31.3	N/A	N/A	34.7	20.7	12.5	8.4
Ncuts(TPAMI'00)	39.9	31.2	42.6	25.4	N/A	N/A	34.8	15.5	12.9	7.4
ALLKM	46.4	31.4	46.9	31.5	N/A	N/A	28.7	11.6	12.8	7.1
AllNcuts(TPAMI'00)	42.8	25.2	43.7	25.5	N/A	N/A	35.2	20.3	14.9	9.4
MvSCN (IJCAI'19)	45.0	35.0	67.7	61.3	N/A	N/A	44.6	35.6	17.5	13.0
EAMC (CVPR'20)	41.9	25.6	38.9	21.4	69.0	75.3	42.2	32.6	16.2	11.9
MVC-VAE (AAAI'20)	39.9	28.1	70.8	58.5	N/A	N/A	52.3	35.2	14.1	11.5
DEMVC (InfoSci'21)	39.4	22.2	38.7	27.0	69.4	77.9	45.1	32.5	15.9	11.9
SiMVC (CVPR'21)	50.8	47.1	56.9	50.4	80.8	89.6	36.8	23.1	14.2	10.4
CoMVC (CVPR'21)	46.6	42.6	54.1	50.4	89.4	89.6	51.4	37.4	16.7	13.0
MFLVC (CVPR'22)	60.6	52.8	63.1	56.6	N/A	N/A	48.5	34.9	16.9	12.8
SPDMC (TNNLS'23)	64.4	50.6	70.1	63.0	N/A	N/A	47.8	31.7	12.9	9.0
ICMVC (AAAI'24)	49.6	37.9	64.7	53.7	81.0	86.6	36.4	30.3	13.3	12.7
DIVIDE (AAAI'24)	64.1	52.9	71.6	58.5	79.0	86.6	31.4	12.4	**19.1**	12.8
Ours	**71.9**	**66.1**	**76.3**	**71.6**	**100**	**100**	**56.9**	**40.1**	18.5	**15.0**

Table 2: Clustering performance on various kinds of datasets (The **bold** denotes the best while underline the second best).

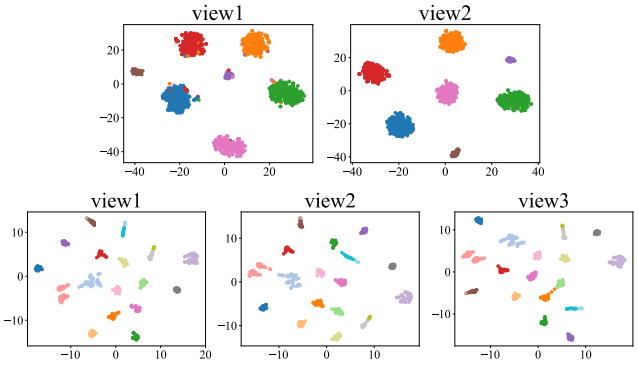

Figure 2: The t-SNE visualization of STCMC-UR on Caltech-2V and COIL20 dataset.

within the same event category. **NUS22**[4] consists of 10,155 images across 22 classes from the original NUS-WIDE-Object dataset, in which two kinds of features are adopted.

Compared Methods. We select 4 traditional Single/All-view clustering methods and 10 latest multi-view clustering methods for comparison to demonstrate the superiority of our method. The classical clustering algorithms are K-Means(KM), Normalized Cuts(Ncuts) and all-view version of them. Other methods include MvSCN (Huang et al. 2019), EAMC (Zhou and Shen 2020), MVC-VAE (Yin, Huang, and Gao 2020), DEMVC (Xu et al. 2021), SiMVC (Trosten et al. 2021), CoMVC (Trosten et al. 2021), MFLVC (Xu et al. 2022), SPDMC (Chen et al. 2023), ICMVC (Chao, Jiang, and Chu 2024), DIVIDE (Lu et al. 2024b).

[4]https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

Implement Details. We ran 20 times and each for 100 epochs on PyTorch 1.13.0 platform (Python 3.8) equipped with a 24GB NVIDIA RTX-4090D GPU on Windows 10 system. We used two widely applied metrics for evaluation: ACC (Accuracy) and NMI (Normalized Mutual Information). Higher values for both metrics indicate better performance. For the datasets, the training batch size was consistently set as 100. The Adam optimizer was adopted with the learning rate of 0.001.

Experimental Results

Table 2 illustrates the clustering performance of our proposed method across five datasets. From the table, we have achieved substantial improvements across all datasets compared with the traditional clustering methods KM, Ncuts, ALLKM and AllNcuts. In addition, the proposed STCMC-UR consistently outperforms other MVC methods in almost all datasets. On the Caltech-2V dataset, our method achieves significant improvement in ACC and NMI compared to the second-best (SPDMC) method, with increases of 7.3% and 15.5%. Our method performs also competitively in large datasets. For example, on NUS22, compared to the SiMVC, CoMVC and MFLVC, STCMC-UR surpassed their results by margins of 3.6%, 1.1% and 0.9%, respectively. In challenging datasets with high complexity, such as COIL20, STCMC-UR demonstrates remarkable performance, achieving 100% clustering results in terms of ACC and NMI. This exceptional performance can likely be attributed to our innovative integration of evidences and the view-weight learning mechanism, which ensure high confidence in the clustering decisions. Additionally, the results of some methods on the COIL20 dataset are N/A because these methods require a feature matrix input, which differs from the original image pixel format used in this paper.

Methods	Caltech-2V		Caltech-3V		COIL20		Event8		NUS22	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
(1) $\mathcal{L}_C$	48.9	48.0	23.4	19.2	86.7	96.1	43.0	28.2	11.0	7.9
(2) $\mathcal{L}_C + \mathcal{L}_{CL}$	63.6	55.3	74.2	67.0	91.0	96.0	48.8	37.4	16.7	13.7
(3) $\mathcal{L}_C + \mathcal{L}_{CE}$	66.4	60.7	27.9	40.5	90.0	97.7	45.2	32.2	12.4	9.9
(4) $\mathcal{L}_C + \mathcal{L}_{CL} + \mathcal{L}_{CE}$	**71.9**	**66.1**	**76.3**	**71.6**	**100**	**100**	**56.9**	**40.1**	**18.5**	**15.0**

Table 3: Ablation study on different multi-view datasets.

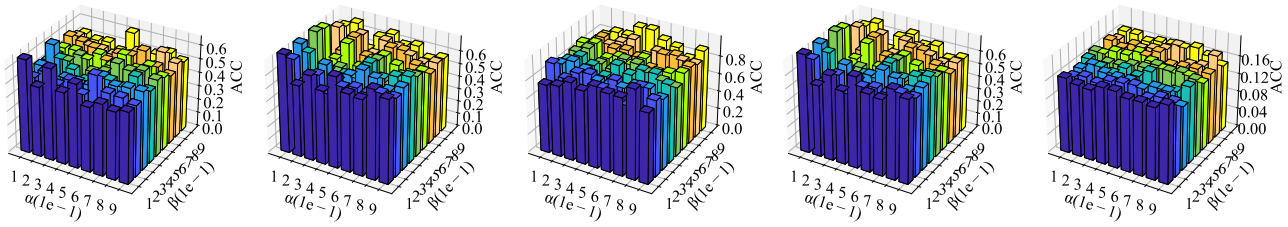

Figure 3: Parameter analysis of STCMC-UR on Caltech-2V, Caltech-3V, COIL20, Event8, and NUS22 dataset.

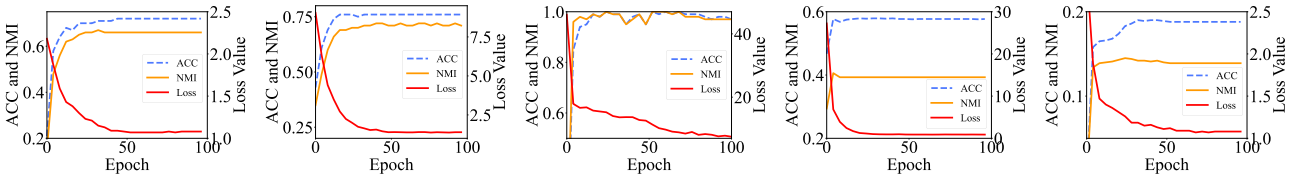

Figure 4: Convergence analysis of STCMC-UR on Caltech-2V, Caltech-3V, COIL20, Event8 and NUS22.

Ablation Study

To validate the effectiveness of each component of our function, we conducted additional ablation experiments. The experimental results in Table 3 indicate that the performance is the worst across all datasets when the loss function includes only $\mathcal{L}_C$, However, adding $\mathcal{L}_{CL}$ and $\mathcal{L}_{CE}$ individually to the $\mathcal{L}_C$ results in noticeable performance improvement. The best performance is reached when all the modules, $\mathcal{L}_C$, $\mathcal{L}_{CL}$ and $\mathcal{L}_{CE}$, are combined in the loss function. Our ablation study reveals that: 1) All of the different modules contribute to the final performance gain; 2) In some datasets with fewer views, such as Caltech-2V, the evidence generation module plays a very important role.

Visualization Clustering Results

To vividly show the clustering quality of our method, we present the t-SNE visualization of the clustering results in Figure 2 for two datasets due to lack of space, Caltech-2V and COIL20. The visualization on both datasets provides compact intra cluster and separable inter cluster structure. It reflects the capability of our method in integrating information from multiple views for clustering.

Parameter Sensitivity Analysis

To assess the parameter sensitivity of our method, we employed a grid search approach, varying α and β within the range (0,1) at intervals of 0.1. The experimental results are displayed in the Figure 3. As observed, our method performs well across all datasets with no significant deterioration in model performance on most parameter settings. This indicates that our method is not sensitive to parameter variations and demonstrates stable performance.

Convergency Analysis

To investigate the convergence of our STCMC-UR, we perform a series of experiments to reveal the objective loss and clustering metrics values of ACC and NMI in Figure 4. It is observed that the loss values decrease rapidly at first, and then it tends to keep a stable value after about 40 epochs. Meanwhile, the clustering metrics also increase and reach an unchanged period after a few epochs. Both of them validate the convergence property of our method.

Conclusion

This paper introduces a STCMC-UR method, which integrates the contrastive learning and uncertainty obtained from trusted learning for clustering. It can fully explore the discriminative feature information with uncertainty refined self-supervised contrastive learning for achieving better clustering performance. However, the proposed method fails to deal with incomplete multi-view data, where missing values in certain views create imbalances in information content across views. We plan to address these limitations in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (project no. 62206254 and 62176239) and China Postdoctoral Science Foundation (project no. 2024T170843 and 2023M743186).

References

Audun, J. 2001. A logic for uncertain probabilities. international journal of uncertainty. *Fuzziness Knowl. Based Syst*, 9(3): 212–279.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *ICML*, 1613–1622. PMLR.

Chao, G.; Jiang, Y.; and Chu, D. 2024. Incomplete Contrastive Multi-View Clustering with High-Confidence Guiding. *AAAI*, 38(10): 11221–11229.

Chen, R.; Tang, Y.; Xie, Y.; Feng, W.; and Zhang, W. 2023. Semisupervised Progressive Representation Learning for Deep Multiview Clustering. *TNNLS*, 1–15.

Chen, Z.; Wu, X.-J.; Xu, T.; Li, H.; and Kittler, J. 2024. Multi-layer multi-level comprehensive learning for deep multi-view clustering. *Information Fusion*, 102785.

Dempster, A. P. 2008. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, 57–72.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 1050–1059. PMLR.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589.

Han, N.; Chen, J.; Zhang, H.; Wang, H.; and Chen, H. 2022a. Adversarial multi-grained embedding network for cross-modal text-video retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–23.

Han, N.; Yang, X.; Lim, E.-P.; Chen, H.; and Sun, Q. 2024. Efficient cross-modal video retrieval with meta-optimized frames. *IEEE Transactions on Multimedia*, 1–14.

Han, Z.; Yang, F.; Huang, J.; Zhang, C.; and Yao, J. 2022b. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *CVPR*, 20707–20717.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022c. Trusted multi-view classification with dynamic evidential fusion. *TPAMI*, 45(2): 2551–2566.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 41–50.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2022. Unsupervised contrastive cross-modal hashing. *TPAMI*, 45(3): 3877–3889.

Hu, S.; Zhang, C.; Zou, G.; Lou, Z.; and Ye, Y. 2024. Deep Multiview Clustering by Pseudo-Label Guided Contrastive Learning and Dual Correlation Learning. *IEEE TNNLS*.

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, Z.; Zhou, J. T.; Peng, X.; Zhang, C.; Zhu, H.; and Lv, J. 2019. Multi-view Spectral Clustering Network. In *IJCAI*, 2563–2569.

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

Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; and Yang, Z. 2019. Deep Adversarial Multi-view Clustering Network. In Kraus, S., ed., *IJCAI*, 2952–2958.

Lin, F.; Bai, B.; Bai, K.; Ren, Y.; Zhao, P.; and Xu, Z. 2022. Contrastive multi-view hyperbolic hierarchical clustering. *arXiv preprint arXiv:2205.02618*.

Lu, Y.; Li, Q.; Zhang, X.; and Gao, Q. 2024a. Deep contrastive representation learning for multi-modal clustering. *Neurocomputing*, 581: 127523.

Lu, Y.; Lin, Y.; Yang, M.; Peng, D.; Hu, P.; and Peng, X. 2024b. Decoupled Contrastive Multi-View Clustering with High-Order Random Walks. *AAAI*, 38(13): 14193–14201.

Mao, Y.; Yan, X.; Guo, Q.; and Ye, Y. 2021. Deep mutual information maximin for cross-modal clustering. In *AAAI*, 8893–8901.

Pedersen, A. B.; Mikkelsen, E. M.; Cronin-Fenton, D.; Kristensen, N. R.; Pham, T. M.; Pedersen, L.; and Petersen, I. 2017. Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 157–166.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shao, Z.; Dou, W.; and Pan, Y. 2024. Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. *Information Fusion*, 103: 102113.

Tao, Z.; Liu, H.; Li, S.; Ding, Z.; and Fu, Y. 2019. Marginalized multiview ensemble clustering. *TNNLS*, 31(2): 600–611.

Trosten, D. J.; Løkse, S.; Jenssen, R.; and Kampffmeyer, M. 2021. Reconsidering Representation Alignment for Multi-View Clustering. In *CVPR*, 1255–1265.

Wang, Q.; Cheng, J.; Gao, Q.; Zhao, G.; and Jiao, L. 2020. Deep multi-view subspace clustering with unified and discriminative learning. *IEEE Transactions on Multimedia*, 23: 3483–3493.

Wang, R.; Li, L.; Tao, X.; Wang, P.; and Liu, P. 2022. Contrastive and attentive graph learning for multi-view clustering. *Information Processing & Management*, 59(4): 102967.

Wu, N.; Jastrzebski, S.; Cho, K.; and Geras, K. J. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, 24043–24055. PMLR.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable conflictive multi-view learning. In *AAAI*, 16129–16137.

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

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level Feature Learning for Contrastive Multi-view Clustering. In *CVPR*.

Yin, M.; Huang, W.; and Gao, J. 2020. Shared Generative Latent Representation Learning for Multi-View Clustering. In *AAAI*, 6688–6695.

Zhao, H.; Des Combes, R. T.; Zhang, K.; and Gordon, G. 2019. On learning invariant representations for domain adaptation. In *International conference on machine learning*, 7523–7532. PMLR.

Zheng, X.; Tang, C.; Wan, Z.; Hu, C.; and Zhang, W. 2023. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, 11381–11389.

Zhou, R.; and Shen, Y. 2020. End-to-End Adversarial-Attention Network for Multi-Modal Clustering. In *CVPR*, 14607–14616.

Zhu, J.; Cui, Y.; Huang, Z.; Li, X.; Liu, L.; Zeng, L.; and Dai, L.-R. 2024. Adaptive Confidence Multi-View Hashing for Multimedia Retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7900–7904. IEEE.