A Novel Information-based Multi-view Representation Learning

Hongdan Wang[1*] and Jian Zhang[2]

[1] Tieling Normal College, Tieling, 112608, China262737, China
[2] Liaoning Institute of Science and Engineering, Jinzhou, 121010, China
[*] Corresponding author. E-mail: HongdanWangTNC@163.com

Received: December 15, 2024; Accepted: January 19, 2025

Multi-view representation learning methods achieve great performance in various domains via fusing complementary and consistent information of views, which have gained great attention. However, there still exist two issues in current methods. They typically assume strict semantic consistency across views to learn representations of multi-view data. Moreover, they lack a unified theoretical framework in mining patterns of multi-view data, making it difficult to gain a deeper understanding of multi-view representation learning. To this end, a new information-based multi-view representation learning within the encoding-decoding architecture is proposed to aggregate complementary and consistent information for mining patterns (IMVRL). It consists of three mutual information objectives for multi-view representation learning, i.e., concentration learning, consistent learning, and comprehensiveness learning. Three objectives work seamlessly in aggregating complementary and consistent information to mine patterns of multi-view data. Finally, a series of extensive tests across three datasets underscore the advantages and efficacy of IMVRL.

Keywords: Mutual information maximization; multi-view representation learning; deep encoding-decoding architecture

http://dx.doi.org/10.6180/jase.202511_28(11).0008

1. Introduction

With the widespread availability of multi-source data, extracting valuable knowledge and patterns from multiview data has become an important research topic [1–3]. In traditional single-view data analysis, valuable information from other perspectives is often neglected, limiting the model generalization. The introduction of multiview learning methods aims to integrate information from different perspectives, enhancing the accuracy and robustness of data analysis. By establishing relationships and complementarity between multiple views, it is possible to more comprehensively explore inherent structures and patterns of samples [4].

One important branch of multiview learning methods is multiview representation learning. Its goal is to extract effective feature representations from among views, contributing to a better learning of the diversity and inherent structure of the data [5]. This method enhances the model's expressive power and generalization ability by establishing shared representations across different views. In practical applications, multiview representation learning is often used for joint representation learning in data types such as images, text, and speech, and is particularly advantageous when dealing with complex, multimodal data.

Currently, the main methods in multiview representation learning can be divided into traditional methods [5–7] and deep learning-based methods [8–10]. The former mainly includes techniques based on linear models and matrix factorization, which aim to find common subspaces between different views or establish mapping relationships between views to achieve information fusion. Typical techniques such as Canonical Correlation Analysis (CCA) and clustering methods in multiview learning rely on simple mathematical models, offering good interpretability and computational efficiency. However, they have limitations when dealing with large-scale data or nonlinear relation-

ships. Deep learning-based methods, on the other hand, have seen rapid development in various domains, particularly with the maturation of deep learning technologies. Neural network-based approaches to multiview representation learning have gradually become mainstream. These deep methods automatically learn the complex relationships and high-dimensional feature representations between different views through end-to-end training. Common deep methods include Multiview Autoencoders, Multiview Convolutional Neural Networks (CNNs), and graph neural network-based multiview representation learning approaches. These methods show great potential in capturing the nonlinear features of data and enhancing the model's expressive power, especially when dealing with complex structured datasets.

Despite significant progress in multi-view representation learning, two key challenges persist. First, many existing methods rely on rigid assumptions about semantic coherence between different views, which may not always hold in practical applications. As a result, these approaches often struggle to handle heterogeneous and complex data patterns, limiting their ability to generalize effectively to more diverse and noisy datasets. Second, there is a notable lack of a unified theoretical framework to comprehensively explain and guide the design of multi-view learning models. While techniques like multi-view autoencoders, multiview convolutional networks, and graph-based methods have shown strong performance in specific tasks, the absence of a consolidated theory makes it difficult to understand why some models outperform others in particular contexts. Without a solid theoretical foundation, it becomes challenging to assess the strengths, weaknesses, and limitations of these approaches, hindering further advancements in the field.

Information theory has gained increasing attention in deep neural network research, as it provides insights into the impressive success of contemporary deep algorithms. This has led to the exploration of multi-view learning through the lens of information theory. Therefore, a new information-based multi-view representation learning based on encoder and decoder architectures is proposed for exploring patterns of data (IMVRL). It consists of three mutual information objectives for multi-view representation learning, i.e., concentration, consistent, and comprehensiveness. Specifically, the concentration objective maximizes mutual information between the data and its representation within each view, ensuring that each view captures its unique, relevant characteristics. This allows the model to retain complementary information across different views. The consistency objective then aligns these representations by maximizing mutual information between the data and its representations across different views, ensuring that shared semantic patterns are captured consistently. Finally, the comprehensiveness objective focuses on the relationship between the fusion representation (the combined representation of all views) and the view-specific representations, balancing complementary and shared information across views. This synergy allows IMVRL to effectively integrate diverse information from multiple views while preserving both unique and common data patterns. The cooperation of these three objectives results in high-quality representations that capture the full spectrum of the data, enabling the model to achieve superior performance even in the presence of noisy or incomplete data. Finally, extensive experiments on four datasets demonstrate that IMVRL establishes a new benchmark about ACC, NMI, and ARI, confirming its superiority as well as effectiveness.

The rest part is organized as follows: Section 2 defines a new information-based multi-view representation learning method based on encoder and decoder architectures. Section 3 verifies the performance on the real multi-view dataset. Finally, Section 4 concludes the whole work.

2. Methodology

As shown in Fig. 1, a new information-based multi-view representation learning based on the encoder and decoder architectures is proposed to learn inherent patterns of multi-view data (IMVRL). It consists of three mutual information objectives for multi-view representation learning, i.e., concentration, Consistent, and comprehensiveness, which are explained in detail in the following.

2.1. Concentration learning

Let multi-view data with N samples and V views denote as $X = \left\{X^{(v)}\right\}_{v=1}^{V}$, where $X^{(v)} = \left\{x_1^{(v)}, x_2^{(v)}, \ldots, x_N^{(v)}\right\}$ denotes N independent and identically distributed data in the t-th view. For multi-view data, each view typically contains unique information about the samples that cannot be provided by other views. To fully capture inherent information of each view, the concentration learning (CL) strategy is designed to learn view-specific representations via maximizing the mutual information between data and representations, which enables the model to fully exploit the potential of each view.

Specifically, given data $X^{(v)} = \left\{x_1^{(v)}, x_2^{(v)}, \ldots, x_N^{(v)}\right\}$ in the v-th view, an encoding-decoding architecture is used to generate corresponding view-specific representations via mapping data from high dimensional space into latent

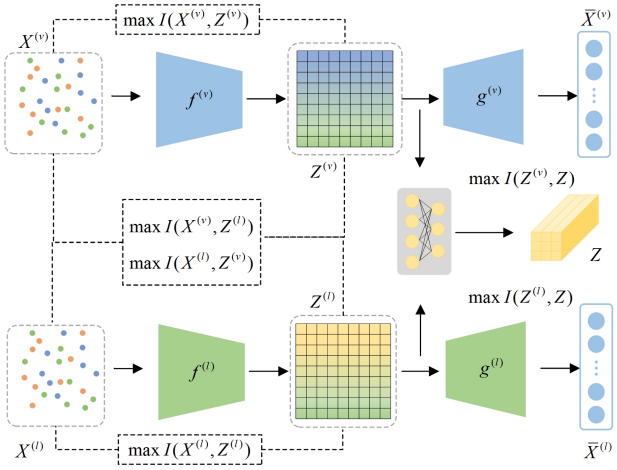


Fig. 1. The illustration of IMVRL. The concentration learning maximizes the mutual information $I\left(Z^{(v)}, X^{(v)}\right)$ to capture complementary information of views. The consistent learning maximizes the mutual information $I\left(Z^{(l)}, X^{(v)}\right)$ to capture sharing semantic information between views. Meanwhile, the comprehensiveness learning maximizes the mutual information $I\left(Z^{(v)}, Z\right)$ to balance complementary and consistent information between views.

space.

$$Z^v = f^v\left(X^v, W_f^v\right) \quad , \quad \bar{X}^v = g^v\left(Z^v, W_g^v\right) \tag{1}$$

where Z^v and $\bar{X}^v$ denote view-specific representations and generation data, respectively. W_f^v and W_g^v denote parameters of the encoder f^v and decoder g^v, respectively. Then, CL maximizes the mutual information between data and representations to optimize mapping correlations between data space and latent space:

$$\begin{aligned} &\max I\left(Z^v, X^v\right) \\ &= \int dx^v dz^v p\left(x^v, z^v\right) \log \frac{p\left(x^v, z^v\right)}{p\left(x^v\right) p\left(z^v\right)} \\ &= \int dx^v dz^v p\left(x^v, z^v\right) \log \frac{p\left(x^v \mid z^v\right)}{p\left(x^v\right)} \end{aligned} \tag{2}$$

where $I(,)$ denotes the mutual information function. In practical applications, the distribution $p\left(x^v \mid z^v\right)$ is often difficult to compute directly. Therefore, a variational approximation distribution $q\left(x^v \mid z^v\right)$ is introduced to approximate $p\left(x^v \mid z^v\right)$. According to the properties of the Kullback-Leibler divergence, it is stated that:

$$\begin{aligned} &KL\left[p\left(x^v \mid z^v\right), q\left(x^v \mid z^v\right)\right] \\ &= \int dx^v p\left(x^v \mid z^v\right) \log \frac{p\left(x^v \mid z^v\right)}{q\left(x^v \mid z^v\right)} \geq 0 \end{aligned} \tag{3}$$

where $I(,)$ denotes the mutual information function. In practical applications, the distribution $p\left(x^v \mid z^v\right)$ is often difficult to compute directly. Therefore, a variational approximation distribution $q\left(x^v \mid z^v\right)$ is introduced to approximate $p\left(x^v \mid z^v\right)$. According to the properties of the Kullback-Leibler divergence, it is stated that:

$$\begin{aligned} &KL\left[p\left(x^v \mid z^v\right), q\left(x^v \mid z^v\right)\right] \\ &= \int dx^v p\left(x^v \mid z^v\right) \log \frac{p\left(x^v \mid z^v\right)}{q\left(x^v \mid z^v\right)} \geq 0 \end{aligned} \tag{4}$$

This implies that the expected value of the logarithm of $p\left(x^v \mid z^v\right)$ is greater than or equal to that of $q\left(x^v \mid z^v\right)$. Consequently, $p\left(x^v \mid z^v\right)$ in the mutual information expression can be replaced with $q\left(x^v \mid z^v\right)$ to obtain a lower bound:

$$I\left(Z^v, X^v\right) \geq \int dx^v dz^v p\left(x^v, z^v\right) \log \frac{q\left(x^v \mid z^v\right)}{p\left(x^v\right)} \tag{5}$$

Further expanding this expression yields:

$$\begin{aligned} I\left(Z^v, X^v\right) &\geq \int dx^v dz^v p\left(x^v, z^v\right) \log \frac{q\left(x^v \mid z^v\right)}{p\left(x^v\right)} \\ &= \int dx^v dz^v p\left(x^v, z^v\right) \log q\left(x^v \mid z^v\right) \\ &- \int dx^v p\left(x^v\right) \log p\left(x^v\right) \\ &= \int dx^v dz^v p\left(x^v, z^v\right) \log q\left(x^v \mid z^v\right) + H\left(X^v\right) \end{aligned} \tag{6}$$

To implement mutual information maximization for optimizing the encoding-decoding architecture in practice, the empirical data distribution $p\left(x^v, z^v\right)$ is utilized to approximate $p\left(x^v, z^v\right) = p\left(x^v\right) p\left(z^v \mid x^v\right)$, and then the loss of the concentration learning is obtained as follows:

$$p\left(x^v, z^v\right) = \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n^v}\left(x^v\right) \delta_{z_n^v}\left(z^v\right) \tag{7}$$

$$\max I\left(Z^v, X^v\right) \approx \frac{1}{N} \sum_{n=1}^{N} \left[\int dz^v p\left(z^v \mid x_n^v\right) \log q\left(x_n^v \mid z^v\right)\right] \tag{8}$$

where $p\left(z^v \mid x_n^v\right)$ denotes the encoding process and $q\left(x_n^v \mid z^v\right)$ denotes the decoding process.

2.2. Consistent learning

Multi-view data describe the same sample from different aspects, sharing the same semantic information. To capture the correlation between views, consistent learning strategy is designed to maximize the mutual information between data and representations from different views, which helps the model to better understand the same sample from different views. The consistent learning strategy is modeled as follows:

Specifically, given the v-th view data X^v and the l-th view representations $Z^{\wedge}\{(l)\}$, the objective of consistent

learning strategy is modeled as follows:

$$\begin{aligned}&\max\left(X^v, Z^l\right)\\&=\iint p\left(z^l \mid x^v\right)\log\frac{p\left(z^l \mid x^v\right)}{p\left(z^l\right)}dx^v dz^l\\&=KL\left(p\left(z^l \mid x^v\right)p\left(x^v\right)\|p\left(z^l\right)p\left(x^v\right)\right)\end{aligned}\tag{9}$$

where $p\left(x^v\right)$ is the distribution of the input samples and $p\left(z^l \mid x^v\right)$ is the cross-view representation distribution. The adversarial strategy can be utilized to optimize representations to have desired statistical information specific to data. KL denotes kullback-leibler divergence and is unbounded. Thus, Jensen-Shannon (JS) Divergence is used to replace KL divergence:

$$\begin{aligned}&\max\left(X^v, Z^l\right)\\&=JS\left(p\left(z^l \mid x^v\right)p\left(x^v\right)\|p\left(z^l\right)p\left(x^v\right)\right)\end{aligned}\tag{10}$$

Based on the variational estimation of JS divergence,

$$\begin{aligned}&JS\left(p\left(z^l \mid x^v\right)p\left(x^v\right)\|p\left(z^l\right)p\left(x^v\right)\right)\\&=E_{x^r\lceil p(x^v)}\left[\log\rho\left(T\left(x^v\right)\right)\right]+E_{x^v\lceil q(x^v)}\left[\log\left(1-\rho\left(T\left(x^v\right)\right)\right)\right]\end{aligned}\tag{11}$$

where $T\left(x^v\right)=\log\frac{2p(x^v)}{p(x^v)+q(x^v)}$. The objective of consistent learning strate can be rewritten as:

$$\begin{aligned}&\max\left(X^v, Z^l\right)\\&=E_{(x^r,z^l)\sqcap p(z^l|x^r)p(x^r)}\left[\log\rho\left(T\left(x^v, z^l\right)\right)\right]\\&+E_{(x^v,z^l)\square p(z^l)p(x^v)}\left[\log\left(1-\rho\left(T\left(x^v, z^l\right)\right)\right)\right]\end{aligned}\tag{12}$$

To implement mutual information maximization for enhancing consistent learning of multi-view data in practice, the negative sample estimation is used, where positive and negative pairs of data are constructed based on representations. $\rho(T(\cdot))$ is a discriminator to distinguish the negative and positive sample pairs for achieving the correlation maximization between views.

2.3. Comprehensiveness learning

Multi-view fusion representations should capture not only the shared semantic information across different views but also the unique, complementary aspects that each view provides. Therefore, Comprehensiveness learning is designed to achieve a comprehensive mining of multi-view information by maximizing the mutual information between the fusion representations and view-specific representations. Such an objective ensures that the model effectively integrates and utilizes information from all available views, enhancing the overall understanding of the sample.

Specifically, given the view-specific representations Z^v, the objective of the comprehensiveness learning is defined as follows:

$$\max I\left(Z^v, Z\right)\tag{13}$$

Similar to Eq. (7), maximizing mutual information $I\left(Z^v, Z\right)$ is equivalent to

$$\min\left\|G^v Z^v - Z\right\|\tag{14}$$

where Z denotes fusion representations of multi-view data. G^v denotes fusion functions of the v-th views.

2.4. The overall objective function

IMVRL utilizes the following objective function L to optimize the overall model via integrating three mutual information objectives into a unified encoding-decoding multi-view representation learning architecture:

$$\begin{aligned}&L=L_1+\alpha L_2+\beta L_3\\&=-\left(\sum_{v=1}^{V}\max I\left(Z^v, X^v\right)\right)+\alpha\sum_{v=1}^{V}\sum_{l\neq v}^{V}\max I\left(Z^l, X^v\right)\\&+\beta\sum_{v=1}^{V}\max I\left(Z^v, Z\right)\end{aligned}\tag{15}$$

where α and β are hyperparameters that are used to balance the learning of complementary and consistent information of multi-view data. IMVRL is updated iteratively using stochastic gradient descent (SGD) as follows:

$$\theta_{t+1}=\theta_t-\eta\nabla_\theta L\left(B_i;\theta_t\right)\tag{16}$$

where θ_t denotes the parameters at iteration t, η is the learning rate, B_i is a randomly selected mini-batch, and $L\left(B_i;\theta_t\right)$ is the loss function evaluated on B_i. Starting with initialized parameters θ_0, the algorithm updates the parameters at each step by computing the gradient of the loss function over the mini-batch B_i and scaling it by the learning rate. This process is repeated epoch by epoch until a predefined stopping criterion, such as a fixed number of epochs, is met. SGD ensures efficient and incremental parameter updates, progressively refining the model to minimize the loss function and improve performance.

Table 1. The detailed statistics of three datasets.

Datasets	Class	Sample	View	Type
Caltech101-20	20	2386	2	image
LandUse-21	21	2100	2	image
Scene-15	15	4485	2	image

IMVRL leverages a multi-objective framework to integrate complementary and consistent information across

Table 2. The clustering results on Caltech101-20 dataset.

Metric	DFMVC	FDAGF	PVC	PIC	Deal	SCM	CVCL	IMVRL
ACC	0.5956	0.6345	0.4491	0.5746	0.5964	0.6378	0.6218	**0.6618**
NMI	0.5624	0.5822	0.5213	0.5702	0.5688	0.5723	0.5712	**0.5882**
PUR	0.6312	0.6357	0.3577	0.3714	0.3577	0.3233	0.5156	**0.6596**

Table 3. The clustering results on LandUse-21 dataset.

Metric	DFMVC	FDAGF	PVC	PIC	Deal	SCM	CVCL	IMVRL
ACC	0.2031	0.2232	0.2222	0.2255	0.2158	0.1958	0.2058	**0.2305**
NMI	0.2221	0.2441	0.2585	0.2614	0.2612	0.2312	0.2412	**0.2733**
PUR	0.0843	0.0944	0.0875	0.0722	0.0755	0.0562	0.0778	**0.0988**

Table 4. The clustering results on Scene-15 dataset.

Metric	DFMVC	FDAGF	PVC	PIC	Deal	SCM	CVCL	IMVRL
ACC	0.3012	0.3111	0.2491	0.2746	0.2596	0.3090	0.3083	**0.3231**
NMI	0.3884	0.3994	0.2213	0.3302	0.3088	0.3723	0.3805	**0.4002**
PUR	0.2059	0.2011	0.1532	0.1654	0.1557	0.1893	0.1998	**0.2250**

multiple views, offering three key advantages: (1) Concentration learning, which maximizes mutual information within individual views to fully capture view-specific details; (2) Consistency learning, which ensures shared semantic information is consistently captured across views; and (3) Comprehensiveness learning, which preserves both shared and unique information by maximizing mutual information between fusion and view-specific representations. Together, these objectives provide robust, holistic data representations, enabling IMVRL to outperform traditional methods in multi-view tasks by delivering nuanced and accurate representations essential for complex data mining and pattern recognition applications.

3. Result

3.1. Setup

Dataset and Metric: Caltech101-20, LandUse-21, and Scene-15 in multi-view scenarios are utilized in the experiments [11, 12]. The detailed dataset statistics are shown in Table 1. ACC, NMI, and ARI are used to validate the performance between methods. Caltech101-20 contains 2386 samples and 20 classes where each sample has two views. LandUse-21 contains 2100 samples and 21 classes where each sample has two views. Scene-15 contains 4485 samples and 15 classes where each sample has two views.

Implementation Details: The IMVRL model is implemented using the PyTorch framework and trained on an NVIDIA GeForce RTX 3090 GPU. The model's hyperparameters are optimized through a series of experiments to achieve the best performance on each dataset, with specific settings for the number of epochs, learning rate, batch size, and hyperparameters that control the relative importance of different objectives. For the Caltech101-20 dataset, the model is trained for 200 epochs with a batch size of 256 , a learning rate η of 0.001 , and the regularization hyperparameters α and β set to 0.05 each. For the LandUse-21 and Scene-15 datasets, the model is trained for 500 epochs with a batch size of 256 , a learning rate of 0.0001 , and both α and β are set to 0.5 . The architecture consists of 4 encoder layers and 3 fusion layers across all datasets. The encoder layers are designed to capture the data's intrinsic features, while the fusion layers combine information across different views, ensuring the model learns comprehensive representations.

Comparison Methods: Seven methods are compared on three datasets about ACC, NMI, and ARI, to demonstrate the performance of IMVRL, containing DFMVC [13], FDAGF [14], PVC [15], PIC [16], SCM [17], Deal [18], and CVCL [19].

3.2. Comparisons with State-of-the-arts

From the results in Tables 2 to 4, IMVRL outperforms all other baselines on the Caltech101-20, LandUse-21, and Scene-15 datasets about ACC, NMI, and ARI. Specifically, IMVRL obtains the highest results across three metrics on three datasets, demonstrating superior performance of IMVRL. This remarkable performance can be attributed to two primary factors. First, IMVRL employs a sophisticated multi-view representation learning approach that integrates three key mutual information objectives: concentration, consistency, and comprehensiveness. The concentration objective maximizes the mutual information between data and representations within each view, enabling

the model to capture complementary information specific to each view. The consistency objective focuses on aligning the representations across different views by maximizing the mutual information between the data and representations in those views, ensuring that shared semantic information is captured. Finally, the comprehensiveness objective balances both complementary and consistent information by maximizing the mutual information between fusion representations and view-specific representations, resulting in a more comprehensive and robust representation of the multi-view data. Second, the model's architecture ensures effective balancing of complementary and consistent information across views, which is critical for improving overall performance. This holistic approach enables IMVRL to outperform traditional methods, which either rely on single-view data or employ less sophisticated multi-view learning strategies.

Additionally, two important insights emerge from the results: (1) Challenges of Baseline Methods with Multi-View Data: Several baseline methods, including DCCA, DCCAE, and PVC, demonstrate limited performance across all datasets. These models struggle to capture the complex relationships between multiple views, as reflected in their lower clustering accuracy and weaker mutual information performance. This underscores the difficulty of effectively modeling multi-view data using these traditional approaches. In contrast, IMVRL overcomes these challenges by integrating complementary and consistent information across views, enabling superior performance. (2) Advantages of Multi-Objective Learning in IMVRL: The results also highlight the effectiveness of IMVRL's unique multi-objective learning approach. Although methods like PIC and Deal show relatively better performance than other baselines, they still lag behind IMVRL across all evaluation metrics. This performance gap indicates that the combination of the concentration, consistency, and comprehensiveness objectives in IMVRL provides a more balanced and comprehensive strategy for learning from multi-view data, leading to improved representation quality and clustering accuracy.

Table 5. Ablation experiments of IMVRL on Caltech101-20.

L_1	L_2	L_2	ACC	NMI	ARI
√			0.4212	0.4056	0.4352
	√		0.5546	0.5198	0.5977
		√	0.5632	0.5211	0.6211
√	√		0.6068	0.5354	0.6424
√		$\sqrt{2}$	0.6332	0.6023	0.6453
	√	√	0.5862	0.5324	0.5844
√	√	√	0.6618	0.5882	0.6596

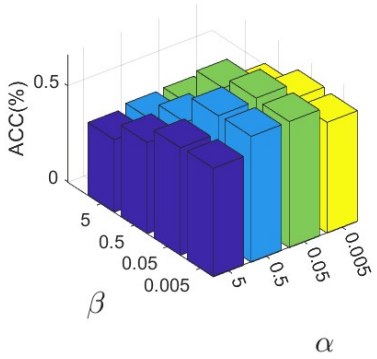
3.3. Ablation Study

IMVRL conducts six ablation experiments on the Caltech101-20 datasets about three metrics to test the effectiveness and design rationality of each loss function. Specifically, (1) IMVRL uses the loss L_1 to train the multi-view encoding-decoding architecture for mining patterns of multi-view data. (2) IMVRL uses the loss L_2 to train the multi-view encoding-decoding architecture for mining patterns of multi-view data. (3) IMVRL uses the loss L_3 to train the multi-view encoding-decoding architecture for mining patterns of multi-view data. (4) IMVRL uses the loss L_2 and the loss L_1 to train the multi-view encoding-decoding architecture for mining patterns of multi-view data. (5) IMVRL uses the loss L_2 and the loss L_3 to train the multi-view encoding-decoding architecture for mining patterns of multi-view data. (6) IMVRL uses the loss L_1 and the loss L_3 to train the multi-view encoding-decoding architecture for mining patterns of multi-view data.

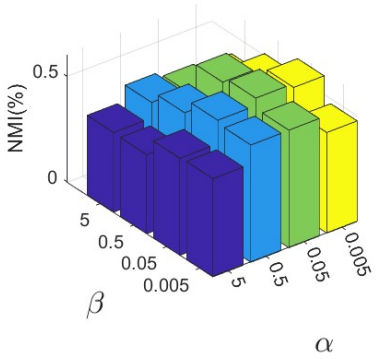
As shown in Table 5, the ablation study reveals three key insights. First, relying on a single loss function (L_1, L_2, or L_3) result in suboptimal performance, as each capture only a specific aspect of multi-view data—view-specific details, shared semantics, or comprehensive balance—failing to model the full complexity. Second, combining two loss functions, such as (L_1 (concentration) and L_2 (consistency), significantly enhances performance by aligning view-specific and shared information, demonstrating the synergistic benefits of multi-objective optimization. Finally, integrating all three loss functions—concentration, consistency, and comprehensiveness—achieves the best results by holistically balancing complementary and shared information, leading to robust and high-quality representations. These findings confirm that IMVRL's multi-objective framework is essential for fully leveraging multi-view data, outperforming single-objective approaches in capturing complex patterns and achieving superior multi-view learning performance.

3.4. Parameter analysis

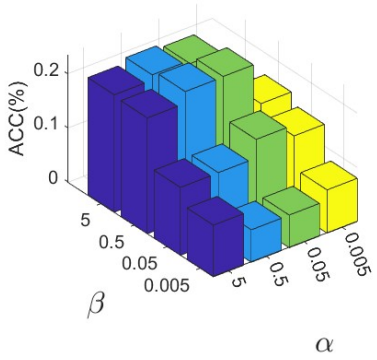
To evaluate the impact of the hyperparameters α and β, which control the balance between the different loss components in IMVRL, we conducted an extensive parameter analysis across three datasets in terms of ACC and NMI. In these experiments, both α and β were varied within a predefined set of values: $\{5, 0.5, 0.05, 0.005\}$. For each combination of these hyperparameters, we recorded the resulting performance and examined how changes in these parameters affect the clustering quality. The experiment results are illustrated in Fig. 2. As observed, IMVRL achieves the best performance on three datasets and three metrics,



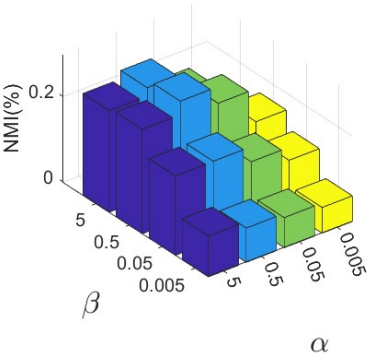
(a) Caltech101-20(ACC)



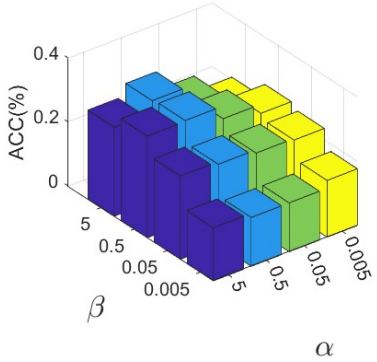
(b) Caltech101-20(NMI)



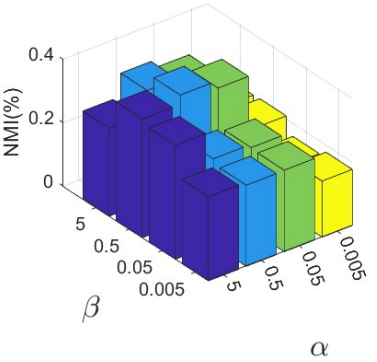
(c) LandUse-21(ACC)



(d) LandUse-21 (NMI)



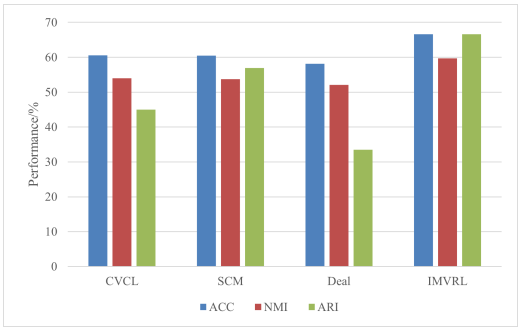
(e) Scene-15 (ACC)



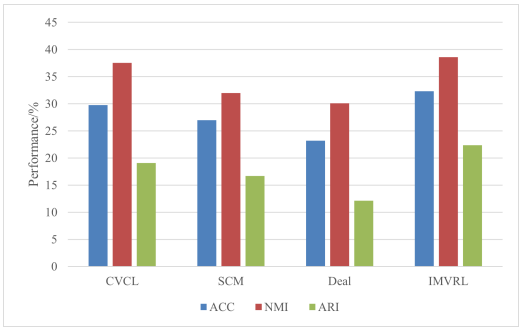
(f) Scene-15 (NMI)

Fig. 2. The hyperparameter analysis of α and β on the three datasets.

when α and β are set within the range 0.05, 0.5. This optimal range indicates that the model exhibits robust performance with respect to the hyperparameters, highlighting their effective role in balancing multi-view complementary learning. Specifically, the value selection of these hyperparameters ensures a suitable trade-off between the different objectives of concentration, consistency, and comprehensiveness, thus improving the overall representation learning and clustering accuracy. Moreover, the experimental results also demonstrate that setting either α and β too high or too low leads to suboptimal performance. For instance, when both hyperparameters are set to values like 5 or 0.005, the clustering quality diminishes, suggesting that overly strong or weak contributions from the individual loss terms hinder the ability to integrate information of views effectively. These findings further underscore the importance of carefully tuning α and β to achieve the best performance, with the chosen optimal values reflecting the inherent robustness of the IMVRL framework in handling multi-view data.



(a) Results on Caltech101-20 of IMVRL.



(b) Results on Scene-15 of IMVRL.

Fig. 3. Robust analysis on the two datasets.

3.5. Parameter analysis

To evaluate the robustness of IMVRL, experiments were conducted by simulating incomplete multi-view data, where 50% of the views were randomly dropped for each sample. This approach tests the model's ability to handle missing information and maintain high performance under challenging conditions. The results, shown in Fig. 3, compare IMVRL with baseline methods on the Caltech101-20 and Scene-15 datasets, using ACC, NMI, and ARI as evaluation metrics. IMVRL consistently outperforms the baselines in all metrics across both datasets, demonstrating its superior robustness in handling incomplete views. The performance gap between IMVRL and the baselines underscores its ability to effectively learn from the available views, ensuring good performance even in the presence of missing or noisy data. This highlights IMVRL's capability to maintain optimal performance in real-world scenarios with incomplete data.

4. Conclusions

IMVRL is an information-based multi-view representation learning method that integrates three core objectives—concentration, consistency, and comprehensiveness—to capture complementary, shared, and balanced information across views. By maximizing mutual information within and across views, IMVRL achieves high-quality representations, consistently outperforming existing methods on benchmark datasets. However, challenges remain in scalability, hyperparameter optimization, generalization to diverse views, and efficient mutual information estimation. Addressing these will be critical for enhancing IMVRL's applicability to real-world scenarios.

References

[1] C. Xu, J. Si, Z. Guan, W. Zhao, Y. Wu, and X. Gao. "Reliable conflictive multi-view learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***38***. 14. 2024, 16129–16137. DOI: 10.1609/aaai.v38i14.29546.

[2] W. Hu, Y. Wu, and Z. Yang, (2024) *"An Analysis of Credit Risk Prediction for Small and Micro Enterprises"* **Journal of Artificial Intelligence Research *1***(2): 1–21. DOI: 10.70891/JAIR.2024.110004.

[3] J. Gao, P. Li, A. A. Laghari, G. Srivastava, T. R. Gadekallu, S. Abbas, and J. Zhang, (2024) *"Incomplete multiview clustering via semidiscrete optimal transport for multimedia data mining in IoT"* **ACM Transactions on Multimedia Computing, Communications and Applications *20***(6): 1–20. DOI: 10.1145/3625548.

[4] L. Fu, S. Huang, L. Zhang, J. Yang, Z. Zheng, C. Zhang, and C. Chen, (2024) *"Subspace-contrastive multi-view clustering"* **ACM Transactions on Knowledge Discovery from Data *18***(9): 1–35. DOI: 10.1145/367483.

[5] Y. Sun, Y. Qin, Y. Li, D. Peng, X. Peng, and P. Hu, (2024) *"Robust multi-view clustering with noisy correspondence"* **IEEE Transactions on Knowledge and Data Engineering**: DOI: 10.1109/TKDE.2024.3423307.

[6] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2023) *"Deep incomplete multi-view clustering via information bottleneck for pattern mining of data in extreme-environment IoT"* **IEEE Internet of Things Journal**: DOI: 10.1109/JIOT.2023.3325272.

[7] S. Shi, F. Nie, R. Wang, and X. Li, (2021) *"Multi-view clustering via nonnegative and orthogonal graph reconstruction"* **IEEE transactions on neural networks and learning systems *34***(1): 201–214. DOI: 10.1109/TNNLS.2021.3093297.

[8] W. Yan, Y. Zhang, C. Lv, C. Tang, G. Yue, L. Liao, and W. Lin. "Gcfagg: Global and cross-view feature aggregation for multi-view clustering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 19863–19872.

[9] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, (2022) *"Dual contrastive prediction for incomplete multi-view representation learning"* **IEEE Transactions on Pattern Analysis and Machine Intelligence *45***(4): 4447–4461. DOI: 10.1109/TPAMI.2022.3197238.

[10] J. Pu, C. Cui, X. Chen, Y. Ren, X. Pu, Z. Hao, S. Y. Philip, and L. He. "Adaptive Feature Imputation with Latent Graph for Deep Incomplete Multi-View Clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***38***. 13. 2024, 14633–14641.

[11] P. Li, A. A. Laghari, M. Rashid, J. Gao, T. R. Gadekallu, A. R. Javed, and S. Yin, (2022) *"A deep multimodal adversarial cycle-consistent network for smart enterprise system"* **IEEE Transactions on Industrial Informatics *19***(1): 693–702. DOI: 10.1109/TII.2022.3197201.

[12] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems *35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[13] P. Zhang, S. Wang, L. Li, C. Zhang, X. Liu, E. Zhu, Z. Liu, L. Zhou, and L. Luo. "Let the data choose: Flexible and diverse anchor graph fusion for scalable multi-view clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***37***. 9. 2023, 11262–11269. DOI: 10.1609/aaai.v37i9.26333.

[14] Y. Ren, J. Pu, C. Cui, Y. Zheng, X. Chen, X. Pu, and L. He. "Dynamic weighted graph fusion for deep multi-view clustering". In: *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*. 2024, 4842–4850.

[15] S.-Y. Li, Y. Jiang, and Z.-H. Zhou. "Partial multi-view clustering". In: *Proceedings of the AAAI conference on artificial intelligence*. ***28***. 1. 2014. DOI: 10.1609/aaai.v28i1.8973.

[16] H. Wang, L. Zong, B. Liu, Y. Yang, and W. Zhou. "Spectral perturbation meets incomplete multi-view data". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2019, 3677–3683.

[17] C. Luo, J. Xu, Y. Ren, J. Ma, and X. Zhu. "Simple Contrastive Multi-View Clustering with Data-Level Fusion". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 2024, 4697–4705.

[18] X. Yang, J. Jiaqi, S. Wang, K. Liang, Y. Liu, Y. Wen, S. Liu, S. Zhou, X. Liu, and E. Zhu. "Dealmvc: Dual contrastive calibration for multi-view clustering". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, 337–346.

[19] J. Chen, H. Mao, W. L. Woo, and X. Peng. "Deep multiview clustering by contrasting cluster assignments". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 16752–16761.