



# Contrastive calibration on consensus and complementary multi-view representations

Negin Jabari <sup>a</sup>, Amjad Seyedi <sup>b</sup>, Reza Mahmoodi <sup>a</sup>, Fardin akhlaghian Tab <sup>a,\*</sup>

<sup>a</sup> Department of Computer Engineering, University of Kurdistan, Iran

<sup>b</sup> Department of Mathematics and Operational Research, University of Mons, Belgium

## ARTICLE INFO

### Keywords:

Data fusion  
Contrastive learning  
Encoder-decoder structure  
Self-representation learning  
Nonnegative matrix factorization

## ABSTRACT

Multi-view representation learning (MRL) aims to exploit information from multiple views to learn discriminative data representations. While most existing methods emphasize consensus learning, they either neglect complementary view-specific information or lack principled mechanisms to balance shared and private representations. Moreover, although recent methods employ increasingly complex architectures to model local structure, they often fail to faithfully preserve the intrinsic structure of the data and lack explicit, strong structural regularization through contrastive objectives that jointly align both intra- and inter-view representations. To address these limitations, we propose C<sup>4</sup>MV, a novel MRL framework that explicitly integrates consensus and complementary representation learning with contrastive calibration. Unlike prior methods, C<sup>4</sup>MV jointly learns shared and view-specific representations through joint and disjoint self-representation factorizations, implemented via coordinated nonnegative matrix factorizations with diversity regularization to prevent redundancy across views. Furthermore, we introduce a contrastive calibration regularization that aligns intra- and inter-view representations using contrastive graph constraints, enhancing sample-level discriminability while reducing reliance on negative pairs. This unified formulation enables balanced fusion of multi-view information and faithful preservation of intrinsic data structure. The resulting optimization problem is solved using an efficient iterative algorithm. Extensive experiments on real-world datasets demonstrate that C<sup>4</sup>MV consistently outperforms state-of-the-art unsupervised multi-view representation learning methods. The source code is publicly available at: <https://github.com/neginjabari/C4MV>.

## 1. Introduction

Multi-view Representation Learning (MRL) has attracted considerable attention due to its ability to incorporate various data from different sources or channels that describe the same entity [1]. Several methodologies have been developed for MRL, including Canonical Correlation Analysis [2], Subspace Clustering [3], Graph-based learning [4], and Nonnegative Matrix Factorization (NMF) [5]. Recent developments in MRL have primarily concentrated on investigating underlying complementary information [6], resolving incomplete multi-view problems, and utilizing tensor learning techniques. In addition, researchers have suggested methods to achieve agreement across different perspectives, conserve energy by including embedded projections, and include both affine and nonnegative restrictions [1]. The purpose of these methods is to systematically record multi-way interactions and reveal the fundamental structure of multi-view data. This helps with subsequent tasks, such as clustering [7], feature selection [8], and object recognition [9].

MRL has garnered considerable interest, with NMF emerging as a potent technique for representing and clustering multi-view data. NMF provides interpretations based on parts and effectively reduces dimensionality [10]. Classic factorization-based multi-view methods enhance performance by utilizing Collective Matrix Factorization (CMF) to jointly factorize matrices with shared entities, thereby learning unified low-rank representations and capturing cross-view structures [11]. NMF-based models that adopt CMF principles yield more robust and informative clustering than traditional single-view approaches [12]. Multi-NMF [5] represents an influential contribution to introducing the concept of cooperative multi-view NMF learning, which aims to unify representations from different viewpoints together to reach a shared agreement. Building upon this foundation, progress in NMF techniques has tackled difficulties in multi-view situations by enforcing orthogonality constraints [13], manifold learning [14], investigating local geometric features [15], and creating consensus manifold representations across views [16]. These approaches aim to effectively identify and

\* Corresponding author.

E-mail addresses: [negin.jabbari@uok.ac.ir](mailto:negin.jabbari@uok.ac.ir) (N. Jabari), [seyedamjad.seyedi@umons.ac.be](mailto:seyedamjad.seyedi@umons.ac.be) (A. Seyedi), [reza.mahmoodi@uok.ac.ir](mailto:reza.mahmoodi@uok.ac.ir) (R. Mahmoodi), [f.akhlaghian@uok.ac.ir](mailto:f.akhlaghian@uok.ac.ir) (F. akhlaghian Tab).

<https://doi.org/10.1016/j.patcog.2026.113291>

Received 9 August 2025; Received in revised form 24 December 2025; Accepted 9 February 2026

Available online 11 February 2026

0031-3203/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

combine information from multiple perspectives while maintaining the integrity of the data's structure and managing data with a high number of dimensions and sparsity. The following methods further developed this idea by integrating tackling imbalances in the dataset [17], introducing auto-weighting schemes [12], and investigating semi-supervised scenarios [18]. More recent advances have focused on leveraging self-representation strategies [19] and deep NMF models [20,21] to capture intricate and hierarchical semantic information better and reduce feature redundancy.

While progress has been made, these consensus-based methods have largely concentrated on learning shared representations, often at the cost of preserving unique, view-specific information. As a result, they tend to overemphasize common structures, which can lead to the inclusion of redundant or less informative features. Moreover, they frequently neglect the heterogeneity across different views, yielding representations that are either redundant or suboptimal [1]. To overcome these limitations, methods such as DiNMF [22] introduce diversity-promoting regularization, which minimizes inner-product correlations between view-specific representation matrices. This encourages orthogonality and reduces feature overlap, promoting the learning of more distinct and informative features across views. Similarly, the NRRNMF-ML method introduces a new non-redundant regularizer utilizing the Hilbert-Schmidt independence criterion (HSIC) [6] to minimize redundant features and encourage distinct features across views. Combining this approach with manifold learning aims to achieve a more comprehensive understanding of multi-view data.

Recent research has acknowledged the significance of concurrently utilizing both consensus and complementary information to achieve efficient representation [23]. The consensus information represents the underlying structure that is shared by different perspectives, while the complementary information captures the unique traits of each perspective [22,24]. Several methodologies, including 2CMV [23], CCNMF [25], and ECNMF [26], have been suggested to extract and integrate these two types of information simultaneously. However, current techniques often face challenges in efficiently extracting and balancing both consensus and complementary information. Despite recent advancements, current techniques often struggle to extract and balance both consensus and complementary information efficiently. There remains significant potential to enhance their discriminatory power, particularly in complex scenarios.

Alongside this line of research, advances in data representation have explored integrating autoencoder principles with NMF [27]. Although traditional autoencoders are effective for unsupervised representation learning [28], they often lack interpretability because they do not account for the nonnegative structure of the input data. To address this, encoder-decoder NMF approaches have been introduced to learn low-dimensional representations by incorporating both encoding and decoding processes, rather than relying solely on decoder-based reconstruction. Approaches such as the Non-negative Symmetric Encoder-Decoder (NSED) [29] aim to capture latent features while preserving data nonnegativity. This representation strategy is conceptually related to transform learning, a complementary representation learning and fusion paradigm that focuses on learning data-adaptive transforms for compact and discriminative feature extraction [30,31]. Related work has further highlighted the limitations of traditional decoder-only NMF methods, which lack refinement or verification mechanisms within a self-representation framework. By jointly learning encoder and decoder factorizations [29], encoder-decoder NMF models capture projected latent features while preserving nonnegativity. This autoencoder-like structure combines the strengths of factorization-based methods with the representation learning and noise tolerance of autoencoders, while reducing information loss during feature extraction, making them suitable for multi-view representation tasks with noisy data. However, despite these strengths, such models [27] remain sensitive to regularization design, as inadequate regularization can lead to degenerate representations that fail to capture the underlying data structure.

Another complementary direction is contrastive learning, which has gained attention for extracting intrinsic supervisory signals from unlabeled data. Its core idea is to increase similarity between positive samples while decreasing similarity among negative ones in the representation space. Early methods such as Noise Contrastive Estimation (NCE) [32] and InfoNCE [33] enabled more advanced frameworks like MoCo [34] and SimCLR [35], which focus on learning image representations through contrastive mechanisms. Contrastive learning has also been applied to multi-view clustering, for example CMC [36], which extracts semantic information, and MVGRL [37], which leverages graph diffusion matrices for multi-view contrastive tasks. However, most multi-view contrastive methods primarily emphasize aligning highly similar samples across views, often overlooking the diverse yet correlated relationships among less similar samples. This narrow focus can limit the ability to capture rich and nuanced structures in complex multi-view data. To address this issue, DealMVC [38] introduces a contrastive calibration mechanism that enforces consistency among similar but distinct samples across views and aligns cross-view feature similarity with high-confidence pseudo-labels. Despite being framed as an unsupervised approach, DealMVC relies on pseudo-labels generated by a trained classifier.

To improve the performance of multi-view representation learning in addressing the aforementioned challenges, this paper proposes a self-representation model that integrates feature-level consensus and complementary representations learning with sample-level discrimination via contrastive learning. By leveraging shared latent (joint) and diversity-aware (disjoint) encoder-decoder factorizations, we unify model-level fusion and late-fusion strategies [39] to capture shared and view-specific information. While this formulation enables the extraction of more informative general and view-specific features, it requires a discriminative mechanism to leverage them fully. To this end, we incorporate contrastive learning with a calibration mechanism to enhance representation alignment across views, significantly improving the model's discriminative capability. By increasing positive pair alignment and reducing reliance on negative pairs, this approach enhances efficiency in multi-view settings. We introduce contrastive graph regularization that, instead of using data augmentation, treats corresponding samples across views as positive pairs. For calibration, similar samples across all views are aligned, while dissimilar ones are pushed apart. This setup leverages self-supervisory signals to enforce similarity and dissimilarity constraints, enabling effective calibrated contrastive learning. To the best of our knowledge, this work explores the unification of consensus and complementary principles within an autoencoder-like factorization architecture, and introduces contrastive regularization into NMF-based multi-view learning. Overall, we present Contrastive Calibration on Consensus and Complementary Multi-View representations ( $C^4MV$ ), a unified framework that integrates joint and disjoint autoencoder-like NMFs with contrastive graph regularization. To solve this non-convex joint factorization problem, we employ an efficient iterative optimization method that updates the variables in an alternating manner. Finally, the experimental results of the proposed model validate the effectiveness and superiority of the  $C^4MV$  framework. The key contributions of this work are summarized as follows:

- The fundamental model is a self-representation multi-view NMF that utilizes encoder and decoder factorizations to extract informative part-based representations across multiple views.
- By integrating joint and disjoint factorizations, which realize model-level and late fusion strategies, the model effectively captures both consensus and complementary multi-view representations.
- This model enhances discrimination representations by applying contrastive calibration regularization, leveraging attractive forces between corresponding samples from different views and repulsive forces to separate unrelated ones.
- The consensus and complementary representations learning, along with contrastive graph regularization, are integrated into a uni-

fied end-to-end framework and solved via an iterative optimization method.

- Comprehensive experiments on nine datasets indicate the excellence of the  $C^4MV$  framework, highlighting its improved clustering performance.

The remainder of this paper is organized as follows. Section 2 reviews related work on multi-view learning and provides background on NMF. Section 3 introduces the proposed model along with the numerical optimization approach. Section 4 presents the experimental results on real-world multi-view data. Finally, the conclusion is described and future work is presented in Section 5.

## 2. Preliminaries

In this section, related work on multi-view learning approaches is reviewed, including both representation and clustering methods. We also discuss approaches that leverage autoencoder-factorization architectures and contrastive learning in the context of multi-view learning. We then introduce the main notations and definitions used throughout this paper for clarity and consistency. Finally, we provide an overview of NMF, a fundamental matrix decomposition technique that underpins our proposed method.

### 2.1. Related work

The NMF has contributed to the growth of multi-view representation, as it is characterized by interpretative simplicity and efficiency in mapping high-dimensional data into low-dimensional, nonnegative representations that capture part-based underlying structure and latent patterns. Foundational studies in this area were led by Liu et al. [5] that proposed a model aligning multiple views through a shared consensus matrix while offering balanced representation among the views. This seminal work paved the path for more complex NMF adaptations, which later aimed to grasp the multi-view data's inherent complexities and structural variations. In multi-NMF, consensus-based techniques strongly emphasize integrating various data views into a common representation. Liang et al. [40] introduced NMF-CC, a factorization-based multi-view clustering model with co-orthogonal constraints. These constraints eliminate redundancy and help ensure the consensus matrix accurately captures shared characteristics across views. By using Triplex Regularized NMF for Re-weighted Multi-view Clustering, Feng et al. [41] expanded on this idea. This approach balances views' contributions according to their significance by dynamically adjusting each view's weight using triplex regularization.

Manifold learning is based on the idea that high-dimensional data lie on a smooth, low-dimensional manifold. This allows dimensionality reduction methods to retain the data's essential local and sometimes global geometric properties. To preserve the local geometric structure of the data and enable a unified representation solution across different views, manifold learning can be integrated with NMF. This combination helps to capture both the intrinsic, potentially non-linear shape of the data and the underlying low-rank structure. For instance, Khan et al. [42] proposed an NMF framework with manifold regularization for multi-view data clustering. By preserving the local geometric structure of the original data, this approach enhances the robustness of the low-dimensional representations against noise and improves their ability to capture the true underlying relationships in the data. Building on this, Multi-manifold Regularized NMF (MMNMF) [14] treats each view as an independent manifold and integrates them into a unified consensus manifold, effectively fusing local geometries from all views. Feng et al. [41] propose SMCTN, which combines graph, pairwise, and consensus regularizations to preserve local geometries and enhance view alignment. Liu et al. [12] extend this idea using graph dual regularization to model both intra-view and inter-view relationships, further refining the alignment and structural consistency across views. Tang et al. [15]

align manifold structures across views using pairwise co-regularization and dual graph regularization, leveraging both data and feature space graphs.

Subsequently, to address the limitations of earlier models in handling view-specific variations, researchers introduced approaches that explicitly extract diverse and complementary information from each view. Wang et al. [22] presented Diverse NMF (DiNMF), which integrates diversity constraints to extract distinct information from each view, hence enhancing the clustering process. This adaptation is a key factor in dealing with data from heterogeneous sources, as it enhances the model's ability to maintain accuracy in representing varied multi-view data characteristics. Li et al. [43] expanded this concept by introducing Robust Multi-view NMF with Adaptive Graph and Diversity Constraints, which integrates adaptive graph learning to reduce inter-view redundancy while maintaining local structures. More recently, to enhance DiNMF, Zhang et al. [44] introduced orthogonal diversity NMF (ODNMF), which enforces basis-matrix orthogonality for cleaner representations, and uses diversity enhancement to enrich the extracted information. Cui et al. [6] proposed Nonredundancy Regularization NMF with Manifold Learning (NRRNMF-ML), which introduces an HSIC-based regularization into the NMF function. This regularizer encourages each view to contribute uniquely while reducing redundancy among representations. Additionally, a manifold regularizer is incorporated to retain the local structure of each view.

Following the exploration of diversity, the need to integrate shared (consensus) information and view-specific (complementary) features has led to further innovations. Luong and Nayak [23] proposed an NMF-based model for learning consensus and complementary information from multi-view data (2CMV). This method integrates manifold learning to maintain the geometric structure of the data. An orthogonality-based enhancement term is proposed to differentiate and isolate the consensus and complementary components inside the representation, maximizing the extraction of coherent and unique information across various viewpoints. Li et al. proposed the Consensus and Complementary Regularized NMF (CCNMF) method [25], which uses NMF with two regularization terms to learn complementary and consensus representations. These representations are fused into an integrated graph for clustering, preserving the data's local structure through graph regularization. Huang et al. propose the ECNMF [26] framework for learning robust and meaningful multi-view data representations. The key idea is to decompose the representation of each view into two parts: a view-specific part and a shared part. The exclusivity term is designed to increase the difference between the specific and shared components in each view. Such a mechanism has encouraged the model to capture unique information from every view well. On the other hand, the consistency term tries to reduce the difference between the shared components of the views so that the common information can be easily extracted from the views.

Alongside this, modern representation learning methods, particularly those based on autoencoders, have advanced MRL by effectively capturing complex underlying data structures. Motivated by autoencoder-like NMF methods, multi-view self-representation has been introduced, enabling the joint learning of reconstruction and consensus embeddings. Xiang et al. [27] proposed the dual auto-weighted multi-view clustering via autoencoder-like NMF ( $DA^2NMF$ ), which integrated the reconstruction ability of autoencoders with the interpretability of NMF. In this framework, a dual auto-weighted mechanism assigns adaptive weights to each view to emphasize more informative views. More recently, Ban et al. [19] developed ADGNMF, an MRL framework that combines encoder-decoder NMF with dual-graph regularizations to jointly preserve data and feature manifold structures, thereby improving representation learning. Furthermore, several deep autoencoder-like factorization models have been introduced for multi-view representation [20,21]. These models incorporate insights from deep autoencoders to extract the hierarchical semantics of multi-view data in a multi-layer approach. Additionally, to capture the intrinsic local structure within each view, manifold regularizers are incorporated to integrate the

representations of the deep structure. Nonetheless, despite their focus on self-expression and architectural differences, these methods converge on a shared consensus representation, ultimately undermining the preservation of view-specific information.

More recently, the integration of contrastive learning into neural network-based MRL has significantly enhanced the capability to discern and align complementary information across disparate views. Zhang et al. [45] proposed Dual-Weighted Contrastive Learning (DWCL), which improves MRL by selectively reinforcing view representations through a Best-Other contrastive mechanism and applying dual weighting to manage cross-view reliability and prevent representation degeneration. Building on these concepts, Huang et al. [46] introduced MFC-ACL, combining attention mechanisms with a Transformer-based contrastive fusion module to enhance feature consistency and clustering performance. Wu et al. [47] applied multi-view contrastive learning to sentiment triplet extraction, enhancing sentiment analysis by refining triplet representations and improving feature alignment across text views. Additionally, Zhu et al. [48] proposed the Trusted Mamba Fusion Network (TMFN), which incorporates a view selection mechanism and an Average-Similarity Contrastive Learning (AsCL) module to filter noise, maintain cluster consistency, and enhance clustering performance.

The rise of contrastive learning has greatly improved multi-view representation by strengthening the correspondence among different views. Yang et al. [38] introduced a dual contrastive calibration framework (DealMVC) to align view-specific similarity graphs with high-confidence pseudo-labels. This approach utilizes both global and local contrastive losses, and hence, the coherence among similar samples from different views is enhanced, facilitating more accurate clustering results. Finally, Dong et al. [49] introduced the Subgraph Propagation and Contrastive Calibration (SPCC) framework for incomplete multi-view clustering that integrates subgraph propagation with contrastive calibration to cover the misalignment of cluster distributions in multi-view settings. Their approach effectively aligns cluster distributions across views by reconstructing a global structural graph and employing contrastive learning. More recently, SparseMVC [50] tackles cross-view sparsity variation in deep multi-view learning by adaptively encoding each view based on its sparsity ratio. It mitigates representation divergence through contrastive, correlation-guided fusion and further enhances cross-view complementarity with a distribution alignment module.

In summary, despite significant advances in multi-view learning, existing methods still suffer from clear limitations. Most NMF-based approaches emphasize either shared or view-specific representations, while manifold- and graph-regularized variants often collapse multiple views into a single consensus space, suppressing complementary information. Deep autoencoder-based and contrastive multi-view methods improve alignment but rely on complex neural architectures and lack interpretability, with contrastive objectives operating only on latent embeddings rather than on explicit factorized representations. Distinct from prior work, our framework integrates contrastive calibration directly into the NMF formulation, enabling contrastive alignment under nonnegativity and low-rank constraints while explicitly disentangling shared and view-specific components. This integration is non-trivial and bridges factorization-based interpretability with contrastive cross-view alignment, offering a unified and simple alternative to existing deep contrastive multi-view models.

## 2.2. Notations

In this work, scalars are denoted by lowercase italic letters (i.e.,  $i$ ,  $j$ ,  $n$ , etc.) vectors by bold lowercase letters (e.g.,  $\mathbf{a}$ ,  $\mathbf{x}$ ), and matrices by bold uppercase letters (e.g.,  $\mathbf{A}$ ,  $\mathbf{B}$ ). For any matrix  $\mathbf{A}$ ,  $\mathbf{a}_i$  shows the  $i$ th column of  $\mathbf{A}$ , and  $A_{ij}$  means the  $(i, j)$ -element of  $\mathbf{A}$ . The transpose of  $\mathbf{A}$  is written as  $\mathbf{A}^\top$ , and its trace is given by  $\text{Tr}(\mathbf{A})$ . The *Frobenius* norm for matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is defined as  $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2} = \sqrt{\text{Tr}(\mathbf{A}^\top \mathbf{A})}$ . Table 1 provides a summary of the variables used throughout this paper.

**Table 1**

List of variables used in the proposed multi-view learning framework.

Variable	Definition
$n$	number of samples
$n_v$	number of views
$d_v$	feature dimension of $v$ th view
$r$	latent factor
$\mathcal{X}$	multi-view dataset
$\mathbf{X}_v$	$v$ th data matrix
$\mathbf{W}_v$	mapping matrix for $v$ th view
$\mathbf{H}_v$	representation matrix for $v$ th view
$\mathbf{H}_c$	consensus representation matrix
$\mathbf{H}_{\text{comp}}$	complementary representation matrix
$\mathbf{H}_{\text{final}}$	final representation matrix
$\mathbf{M}_v$	similarity matrix for $v$ th view
$\mathbf{C}^+$	positive contrastive graph
$\mathbf{C}^-$	negative contrastive graph
$\mathbf{A}$	diagonal degree graph of $\mathbf{C}^+$
$\mathbf{L}_c$	Laplacian matrix of $\mathbf{C}^+$
$\lambda_1$	diversity control parameter
$\lambda_2$	attractive parameter
$\lambda_3$	repulsive parameter
$\alpha$	consensus-complementary trade-off
$\Theta, \Psi, \Omega$	Lagrangian multiplier matrices

## 2.3. Nonnegative matrix factorization

NMF is one of the popular methods for learning parts-based, low-dimensional representations of nonnegative data [10]. It is widely used for various tasks such as dimensionality reduction [51], clustering [52], and link prediction [53]. Let  $\mathbf{x} \in \mathbb{R}_{\geq 0}^d$  be a  $d$ -dimensional vector with nonnegative components, and suppose we have  $n$  such observations denoted by  $\mathbf{x}_j$  for  $j = \{1, 2, \dots, n\}$ . These observations form the input matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}_{\geq 0}^{d \times n}$ . NMF aims to approximate  $\mathbf{X}$  by decomposing it into two nonnegative matrices: a basis matrix  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r] \in \mathbb{R}_{\geq 0}^{d \times r}$  and a coefficient matrix  $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n] \in \mathbb{R}_{\geq 0}^{r \times n}$ , such that  $\mathbf{X} \approx \mathbf{W}\mathbf{H}$  [10]. At the level of an individual sample  $\mathbf{x}_i \in \mathbb{R}_{\geq 0}^d$ , NMF decomposes it into basis  $\mathbf{W} \in \mathbb{R}_{\geq 0}^{d \times r}$  and the representation  $\mathbf{h}_i \in \mathbb{R}_{\geq 0}^r$ . The overall optimization objective for all samples is given by,

$$\min_{\mathbf{W}, \mathbf{H} \geq 0} \sum_i z(\mathbf{x}_i, \mathbf{W}\mathbf{h}_i), \quad (1)$$

where  $z(\cdot)$  is a per-sample reconstruction loss. It is clear that  $\mathbf{h}_i$  represents the weight coefficients used to reconstruct the observation  $\mathbf{x}_i$  as a linear combination of the latent basis vectors in  $\mathbf{W}$ . Each data point is thus approximated as a composition of these basis vectors. Since the number of basis vectors is constrained by the condition  $r \ll \min(d, n)$ , the learned bases are typically incomplete to the original data space. In other words, a close approximation is only achievable if the basis matrix  $\mathbf{W}$  captures the intrinsic structure of the data, as NMF seeks to represent high-dimensional patterns using significantly fewer components. The most common loss function for NMF is the squared Frobenius norm, defined as:

$$\min_{\mathbf{W}, \mathbf{H} \geq 0} z_2(\mathbf{X}, \mathbf{W}\mathbf{H}) = \sum_i \|\mathbf{x}_i - \mathbf{W}\mathbf{h}_i\|^2 = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2, \quad (2)$$

where  $z_2(\cdot)$  represents the least squares loss function.

## 3. Proposed model

This section introduces the proposed C<sup>4</sup>MV (Contrastive Calibration on Consensus and Complementary Multi-View representations) model, designed to effectively capture both shared and distinct patterns across multiple data views. The model integrates three key components: (1) a dual-branch autoencoder-like NMF architecture that jointly learns consensus (shared) and complementary (view-specific) representations; (2)

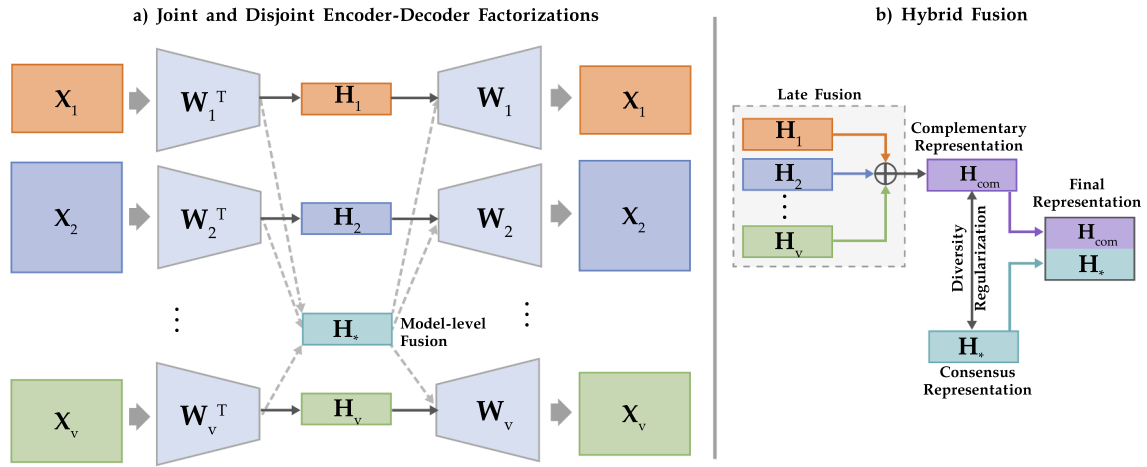


Fig. 1. Overview of the data fusion process in the proposed framework: a) Joint and disjoint encoder-decoders for shared or view-specific feature extraction; b) Hybrid fusion combining model-level and late fusions for richer integration.

a fusion mechanism that unifies these representations into a comprehensive low-dimensional embedding; and (3) a contrastive graph-based regularization to enhance cross-view consistency and inter-view discriminability. To optimize the proposed non-convex objective, we develop an alternating update strategy based on multiplicative update rules.

### 3.1. Consensus and complementary self-representation

Collective Matrix Factorization (CMF) is a data fusion technique that performs a set of joint or disjoint matrix factorizations to extract representations from multiple related data sources. CMFs capture shared or view-specific information, making them well-suited for multi-view representation learning. In this subsection, we introduce a hybrid fusion approach that combines joint (model-level fusion) and disjoint (late-fusion) CMF functions to extract both consensus and complementary representations from multi-view data. These components are cohesively integrated into a unified framework, as depicted in Fig. 1.

#### 3.1.1. Joint autoencoder-like factorization

To capture common features or latent factors shared across multiple data views, we employ a joint factorization approach that learns a consensus representation using a model-level fusion strategy. To implement this and leverage the strengths of self-representation models, we propose a joint autoencoder-like factorization framework. This framework aims to minimize the discrepancy between the original data  $X_v$  for each view  $v$ , and the approximation of these data based on a consensus representation  $H_*$  shared across all views. Formally, the objective function of the model-level fusion is defined as,

$$\min_{W_v, H_*} C_{\text{cons}} = \sum_{v=1}^{n_v} \|X_v - W_v H_*\|_F^2 + \|H_* - W_v^T X_v\|_F^2, \quad \text{s.t. } (W_v, H_*) \geq 0. \quad (3)$$

The first term measures the reconstruction error between the data  $X_v$  and its approximation  $W_v H_*$ , while the second term accounts for the consistency between  $H_*$  and the projections of the data on the basis  $W_v$ . Within our framework,  $H_*$  is defined in  $\mathbb{R}_+^{r \times n}$ , representing the consensus low-rank representation shared across all views, serving as a model-level fusion of multi-view information.

#### 3.1.2. Disjoint autoencoder-like factorization

While the joint factorization systematically extracts consensus information and leverages shared features across all views, it is equally important to ensure that each view's representation preserves and highlights its unique characteristics. This complementary perspective

enriches the overall representation and enhances the model's ability to distinguish among the diverse types of information provided by each view, thereby improving the effectiveness of multi-view data modeling. To explicitly capture view-specific information in a late-fusion paradigm, we introduce the following disjoint autoencoder-like factorization term,

$$\min_{H_v, H_{\text{com}}} C_{\text{comp}} = \sum_{v=1}^{n_v} \|X_v - W_v H_v\|_F^2 + \|H_v - W_v^T X_v\|_F^2, \quad \text{s.t. } (W_v, H_v) \geq 0. \quad (4)$$

In this formulation,  $H_v$  serves as the low-rank representation corresponding to the basis  $W_v$  of the  $v$ th view, following the principles of NMF. This representation is learned directly from each specific view  $v$ , ensuring that  $H_v$  embodies the distinct information inherent in that view, aligning with the concept of semi-coupled transform learning [30].

Furthermore, we define the complementary component matrix  $H_{\text{com}}$  which is the average of the view-specific low-rank representations, i.e.,  $H_{\text{com}} = \sum_{v=1}^{n_v} H_v / n_v$ . This composite matrix,  $H_{\text{com}}$ , not only preserves but enhances the distinctiveness of each view's data, contributing significantly to the model's capability to process and analyze multi-view information comprehensively.

#### 3.1.3. Integrating joint and disjoint factorizations

To learn more informative representations from the data, we introduce a basic objective function by incorporating both consensus and complementary information terms. The objective is to augment the distinction between the latent representations of the consensus and complementary matrices to maximize the extraction of unique information from each data sample. To this end, we propose that these two vectors,  $h_*^{(i)}$  and  $h_{\text{com}}^{(i)}$ , should ideally be orthogonal in low-dimensional space. Orthogonality here is operationalized by minimizing the inner product of the two vectors, which effectively enhances the independence of the encoded features as follows,

$$\min_{H_v, H_*} \mathcal{R}_{\text{div}} = \text{Tr}(H_*^T H_{\text{com}}) = \text{Tr}(H_*^T \sum_{v=1}^{n_v} H_v / n_v), \quad \text{s.t. } (H_v, H_*) \geq 0. \quad (5)$$

Minimizing  $\text{Tr}(H_*^T H_{\text{com}})$  encourages orthogonality between  $H_*$  and  $H_{\text{com}}$ . In this setting, the inner product (and thus its trace) measures second-order dependence, i.e., correlation, between the two representation subspaces, so the regularizer is minimized when their alignment is small and the cross-covariance between the code sets approaches zero. Under standard linear assumptions, zero cross-covariance implies statistical independence, meaning that driving this trace toward zero reduces redundant shared information between the consensus and complementary factors [23]. Concretely,  $H_*$ , obtained from common features across views via the encoder-decoder factorization, is designed to

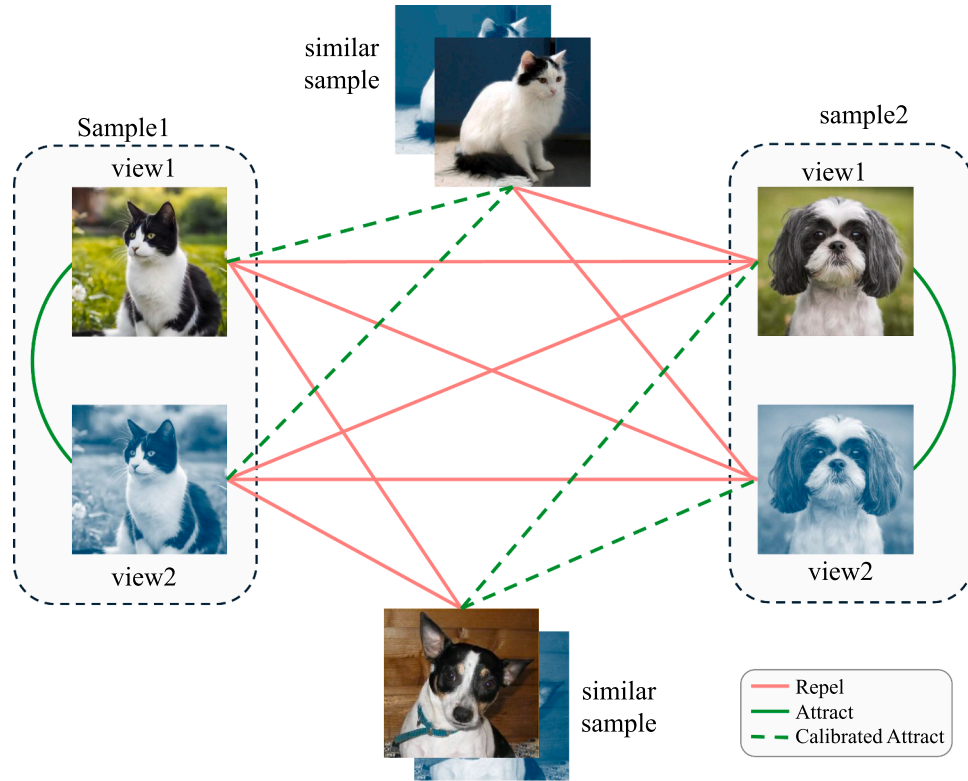


Fig. 2. Illustration of the motivation: Most multi-view representation methods align identical samples across views but overlook consistency between similar samples in different views.

capture the most pervasive shared information. The orthogonality (diversity) constraint then pushes  $H_{\text{com}}$  away from this shared subspace, encouraging it to retain the maximum feasible distinctive, view-specific information. This decorrelation principle mirrors multi-view learning objectives, where enforcing uncorrelated components yields a more non-overlapping decomposition of information, improving identifiability and ensuring each factor explains distinct variations in the data while enriching the overall representation. To unify these considerations into a cohesive objective, we define a joint optimization problem that simultaneously minimizes reconstruction losses and enforces the diversity constraint as follows,

$$\begin{aligned} \min_{W_v, H_v, H_*} & \sum_{v=1}^{n_v} \|X_v - W_v H_*\|_F^2 + \|H_* - W_v^T X_v\|_F^2 \\ & + \|X_v - W_v H_v\|_F^2 + \|H_v - W_v^T X_v\|_F^2 \\ & + \lambda_1 \text{Tr}(H_*^T \sum_{v=1}^{n_v} H_v / n_v), \quad \text{s.t. } (W_v, H_v, H_*) \geq 0, \end{aligned} \quad (6)$$

where  $\lambda_1$  is the diversity control parameter. Although joint and disjoint factorizations capture shared and view-specific structures, the current approach focuses on feature-level representations and may miss important sample-level relationships. In multi-view settings, modeling these interactions is crucial, as they carry rich semantic and structural information not fully captured by reconstruction loss. To overcome this, regularization is needed to guide learning toward more meaningful, discriminative, and non-degenerate embeddings that reflect the data's intrinsic structure.

### 3.2. Contrastive regularization

Contrastive regularization aims to guide the model toward learning representations that emphasize meaningful differences between specific substructures or conditions within the data. There are a limited number of contrastively regularized NMF models that have shown improvements in other tasks, such as single-view data representation [54] and

clustering [52], but their application in multi-view learning remains relatively underexplored. Building on these advancements, we extend contrastive regularization to the multi-view setting, where aligning and distinguishing representations across different views is critical. In this method, contrastive calibration is pivotal in aligning the representations derived from multiple views, effectively calibrating and differentiating between similar and dissimilar data points. Fundamental to our method is contrastive regularization, which is inspired by deep learning techniques that leverage natural data structures for enhanced feature discrimination. We adapt these principles to refine our multi-view data representations, integrating them into our model to ensure robust and discriminative learning outcomes.

#### 3.2.1. Graph construction and pair calibration

In multi-view representation learning, effectively distinguishing between similar and dissimilar samples is critical for robust feature alignment across views. To this end, we introduce a calibration mechanism that refines negative pair selection and enhances sample separation. Fig. 2 illustrates the core motivation behind our contrastive calibration strategy. While traditional multi-view models primarily align identical samples across views, they often overlook relationships between similar, but non-identical, samples. Our method addresses this gap by explicitly modeling both intra-class similarity and inter-class distinction. To calibrate the negative pairs and sharpen the distinction between sample representations in multi-view settings, we first construct similarity matrices  $M^{(v)}$  for each view  $v$ . These matrices serve to identify neighboring samples within the feature space, defining neighbors as those samples  $x_i^{(v)}$  and  $x_j^{(v)}$  that fall within a predefined neighborhood  $\mathcal{N}_k$ ,

$$M_{ij}^{(v)} = \begin{cases} 1, & \text{if } x_j^{(v)} \in \mathcal{N}_k(x_i^{(v)}) \text{ or } x_i^{(v)} \in \mathcal{N}_k(x_j^{(v)}), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Subsequently, we define positive and negative connectivity matrices across all views. The positive connectivity matrix  $C^+$  is computed as

the intersection of the adjacency matrices from all views, thereby ensuring that only those instance pairs consistently deemed similar across different views are treated as positive. This promotes robust alignment by reinforcing similarity among semantically consistent pairs observed across views, as defined below,

$$C_{ij}^+ = \begin{cases} 1, & \text{if } i = j, \\ \min\{M_{ij}^{(v)}\}_{v=1}^{n_v}, & \text{otherwise.} \end{cases} \quad (8)$$

To mitigate the sparsity induced by the minimum operation, we adopt relatively large  $k$  values for both consensus and view-specific neighbor graphs, ensuring that retained positive pairs are semantically consistent across views while providing sufficient density for reliable contrastive supervision. Conversely, the negative connectivity  $C^-$  is derived from the absence of positive connections, establishing a clear distinction for non-neighboring pairs,

$$C_{ij}^- = \begin{cases} \max\{\|\mathbf{x}_i^{(v)} - \mathbf{x}_j^{(v)}\|^2\}_{v=1}^{n_v}, & \text{if } C_{ij}^+ = 0, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

This structure is theoretically motivated as a conservative, worst-case strategy across views that prioritizes global separability over average-case behavior, analogous to max-margin principles in statistical learning theory [32]. To enforce the calibrated relationships between sample pairs, we incorporate both attractive and repulsive regularization terms. These terms are designed to draw similar pairs closer and push dissimilar pairs apart, as detailed in the following sections.

### 3.2.2. Attractive regularization

In the realm of contrastive learning for multi-view data representation, the attractive regularization prioritizes the minimization of distances between positive pairs to emphasize similarity across various representations. This is encapsulated in our attractive regularization term, as formulated below,

$$\begin{aligned} \min_{H_v} \mathcal{R}_{\text{att}} &= \frac{1}{2} \sum_{u,v} \sum_{i,j} \|\mathbf{h}_i^{(u)} - \mathbf{h}_j^{(v)}\|^2 C_{ij}^+ = \sum_{u,v} \text{Tr}(H_u A H_v^T) - \text{Tr}(H_u C^+ H_v^T) \\ &= \sum_{u,v} \text{Tr}(H_u L_c^+ H_v^T), \quad \text{s.t. } (H_u, H_v) \geq 0, \end{aligned} \quad (10)$$

where  $L_c^+$  is the Laplacian matrix corresponding to  $C^+$ , and  $A_{ij} = \sum_{j=1}^n C_{ij}^+$ . This mechanism enhances the alignment of semantically similar instances across views, thereby fostering inter-view consistency and facilitating coherent multi-view integration.

### 3.2.3. Repulsive regularization

Conversely, the repulsive regularization aims to maximize the distance between negative pairs, thereby strengthening the discriminative power of the representations. This mechanism is pivotal in ensuring that dissimilar data points across views remain well-separated. We define the repulsive regularization as follows,

$$\min_{H_v} \mathcal{R}_{\text{rep}} = \sum_{u,v} \sum_{i,j} \mathbf{h}_i^{(u)T} \mathbf{h}_j^{(v)} C_{ij}^- = \sum_{u,v} \text{Tr}(H_u C^- H_v^T), \quad \text{s.t. } (H_u, H_v) \geq 0. \quad (11)$$

In line with the provided framework, we utilize the inner product of low-dimensional representations to assess their similarities, alongside a repulsive term based on the negative pairs, ensuring that representations of dissimilar instances are significantly divergent. The overall contrastive calibration combines the attractive and repulsive terms, moderated by parameters  $\lambda_2$  and  $\lambda_3$  as,

$$\min_{H_v} \mathcal{R}_{\text{contr}} = \lambda_2 \mathcal{R}_{\text{att}} + \lambda_3 \mathcal{R}_{\text{rep}} = \sum_{u,v} \text{Tr}(H_u (\lambda_2 L_c^+ + \lambda_3 C^-) H_v^T), \quad (12)$$

This joint regularization promotes intra-view consistency and enforces clear separation between dissimilar instances across views, thereby enhancing the robustness and discriminative capacity of the learned multi-view representations.

### 3.3. Unified objective function

By integrating the above components into a unified objective function, we define a weighted optimization framework consisting of four terms: the joint factorization  $C_{\text{cons}}$  for cross-view consensus, the disjoint factorization  $C_{\text{comp}}$  for view-specific complementarity, the orthogonal regularization  $\mathcal{R}_{\text{div}}$  for diversity, and the graph contrastive regularization  $\mathcal{R}_{\text{contr}}$  for contrastive calibration concept. The resulting  $C^4$ MV model is formulated as follows:

$$\begin{aligned} \min_{W_v, H_v, H_*} \sum_{v=1}^{n_v} \|X_v - W_v H_*\|_F^2 + \|H_* - W_v^T X_v\|_F^2 \\ + \|X_v - W_v H_v\|_F^2 + \|H_v - W_v^T X_v\|_F^2 + \lambda_1 \text{Tr}(H_*^T \sum_{v=1}^{n_v} H_v / n_v) \\ + \sum_{u,v} \text{Tr}(H_u (\lambda_2 L_c^+ + \lambda_3 C^-) H_v^T), \quad \text{s.t. } (W_v, H_v, H_*) \geq 0. \end{aligned} \quad (13)$$

In the hybrid fusion approach, we combine both model-level and late fusion to combine consensus and complementary representations. Specifically, we define the final representation matrix as,

$$H_{\text{final}} = [\alpha H_*; (1 - \alpha) H_{\text{com}}], \quad (14)$$

where  $H_*$  is the consensus representation learned through model-level fusion and  $H_{\text{com}} = \sum_{v=1}^{n_v} H_v / n_v$  is the averaged complementary representation obtained via late fusion across the views. The parameter  $\alpha \in [0, 1]$  allows for tuning the influence of consensus versus unique view-specific details. This combined approach ensures that  $H_{\text{final}}$  effectively represents both shared and unique data characteristics, enhancing clustering performance by offering a nuanced view of the underlying data structure.

### 3.4. Optimization

This section presents the optimization strategy for the proposed non-convex model. Given that problem (13) involves several variables, we adopt an alternating optimization approach, where each variable is updated sequentially while keeping the others fixed. The specific update rules for each variable are derived as outlined below:

- *Updating rule for the basis matrix  $W_v$*

By keeping all matrices constant except for  $W_v$ , the objective function (13) becomes:

$$\begin{aligned} \min_{W_v} \sum_{v=1}^{n_v} \|X_v - W_v H_*\|_F^2 + \|H_* - W_v^T X_v\|_F^2 \\ + \|X_v - W_v H_v\|_F^2 + \|H_v - W_v^T X_v\|_F^2, \quad \text{s.t. } W_v \geq 0. \end{aligned} \quad (15)$$

To solve the objective (15), a Lagrange multiplier matrix  $\Theta_v$  is introduced to enforce the nonnegative condition on  $W_v$ , yielding the following equivalent objective:

$$\begin{aligned} \min_{W_v, \Theta_v} \mathcal{L}(W_v, \Theta_v) &= \text{Tr}(-2X_v H_*^T W_v^T + W_v H_* H_*^T W_v^T) \\ &+ \text{Tr}(-2H_* X_v^T W_v + W_v^T X_v X_v^T W_v) \\ &+ \text{Tr}(-2X_v H_v^T W_v^T + W_v H_v H_v^T W_v^T) \\ &+ \text{Tr}(-2H_v X_v^T W_v + W_v^T X_v X_v^T W_v) - \text{Tr}(W_v \Theta_v^T). \end{aligned} \quad (16)$$

Taking the partial derivative of  $\mathcal{L}(W_v, \Theta_v)$  with respect to  $W_v$  and setting it to zero yields:

$$\Theta_v = -4X_v (H_v^T + H_*^T) + 2W_v (H_v H_v^T + H_* H_*^T) + 4X_v X_v^T W_v. \quad (17)$$

Applying the complementary slackness condition from the Karush-Kuhn-Tucker (KKT) framework [10], we derive:

$$\Theta_v \odot W_v = 0, \quad (18)$$

where  $\odot$  represents the element-wise (Hadamard) product. Eq. (18) serves as a fixed-point condition that the solution must fulfill upon convergence. By resolving this equation, we obtain the following update rule for  $W_v$ :

$$W_v \leftarrow W_v \odot \frac{2X_v (H_v^T + H_*^T)}{W_v (H_v H_v^T + H_* H_*^T) + 2X_v X_v^T W_v}. \quad (19)$$

- *Updating rule for the complementary matrix  $H_v$*

To update the matrix  $H_v$  while fixing all other matrices, we consider the following objective function in Eq. (13), which is a part of the optimization problem,

$$\begin{aligned} \min_{H_v} & \|X_v - W_v H_v\|_F^2 + \|H_v - W_v^T X_v\|_F^2 + \lambda_1 \text{Tr}(H_v^T H_v / n_v) \\ & + \sum_u \text{Tr}(H_u (\lambda_2 L_c^+ + \lambda_3 C^-) H_v^T), \quad \text{s.t. } H_v \geq 0. \end{aligned} \quad (20)$$

To address this constrained minimization problem (20), we introduce a Lagrangian multiplier matrix  $\Psi_v$ . The Lagrangian  $\mathcal{L}(H_v, \Psi_v)$  integrates the objective function with the non-negativity constraint,

$$\begin{aligned} \min_{H_v, \Psi_v} \mathcal{L}(H_v, \Psi_v) = & \text{Tr}(-2X_v H_v^T W_v^T + W_v H_v H_v^T W_v^T) \\ & + \text{Tr}(H_v H_v^T - 2H_v X_v^T W_v) + \lambda_1 \text{Tr}(H_v^T H_v / n_v) \\ & + \sum_u \text{Tr}(H_u (\lambda_2 L_c^+ + \lambda_3 C^-) H_v^T) - \text{Tr}(H_v \Psi_v^T). \end{aligned} \quad (21)$$

We solve this constrained problem following the same steps as for  $W_v$ : formulating the Lagrangian, applying KKT conditions. Solving this, we update  $H_v$  using the following iterative rule, ensuring non-negativity and optimality,

$$H_v \leftarrow H_v \odot \frac{2W_v^T X_v + \lambda_2 \sum_{u=1}^{n_v} H_u C^+}{W_v^T W_v H_v + H_v + \frac{\lambda_1}{2n_v} H_* + \sum_{u=1}^{n_v} H_u (\lambda_2 A + \lambda_3 C^-)}. \quad (22)$$

- *Updating rule for the consensus matrix  $H_*$*

By fixing all the matrices (except  $H_*$ ), the objective function in Eq. (13) is,

$$\min_{H_*} \sum_{v=1}^{n_v} \|X_v - W_v H_*\|_F^2 + \|H_* - W_v^T X_v\|_F^2 + \lambda_1 \text{Tr}(H_*^T \sum_{v=1}^{n_v} H_v / n_v), \quad \text{s.t. } H_* \geq 0. \quad (23)$$

To manage the constraints and optimize the function (23), we introduce a Lagrangian multiplier matrix  $\Omega$ . The Lagrangian  $\mathcal{L}(H_*, \Omega)$  is then defined as,

$$\begin{aligned} \min_{H_*, \Omega} \mathcal{L}(H_*, \Omega) = & \sum_{v=1}^{n_v} \text{Tr}(-2X_v H_*^T W_v^T + W_v H_* H_*^T W_v^T) + \text{Tr}(H_* H_*^T - 2H_* X_v^T W_v) \\ & + \lambda_1 \text{Tr}(H_*^T \sum_{v=1}^{n_v} H_v / n_v) - \text{Tr}(H_* \Omega^T). \end{aligned} \quad (24)$$

To manage the constraints and optimize the function, we adopt the same strategy used in the update rule for  $W_v$ . Solving the KKT conditions provides us with the update rule for  $H_*$ ,

$$H_* \leftarrow H_* \odot \frac{2 \sum_{v=1}^{n_v} W_v^T X_v}{\sum_{v=1}^{n_v} W_v^T W_v H_* + \frac{\lambda_1}{2n_v} H_v + H_*}. \quad (25)$$

The complete optimization procedure for solving problem (13) using alternating minimization is summarized in Algorithm 1.

### 3.5. Convergence of $C^4MV$

We analyze the convergence properties of  $C^4MV$  under its iterative optimization scheme. The convergence can be established by examining the update rules for  $H_*$ ,  $H_v$ , and  $W_v$ . Without loss of generality, we focus on the update rule of  $H_v$  and prove convergence using the auxiliary function approach [55]. The convergence of  $C^4MV$  with respect to the updates of  $H_*$  and  $W_v$  can be demonstrated in an analogous manner.

**Definition 1.**  $Z(h, h')$  is an auxiliary function for  $F(h)$  if the conditions  $Z(h, h') \geq F(h)$  and  $Z(h, h) = F(h)$  are satisfied [55].

**Lemma 1.** If  $Z$  is an auxiliary function of  $F$ , then  $F$  is non-increasing under the update  $h^{t+1} = \arg \min_h Z(h, h^t)$  [55].

**Proof.**  $F(h^{t+1}) \leq Z(h^{t+1}, h^t) \leq Z(h^t, h^t) = F(h^t)$ .

**Lemma 2.** For nonnegative matrices  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{k \times k}$  (both symmetric) and matrices  $S, S' \in \mathbb{R}^{n \times k}$ , the inequality  $\sum_{i=1}^n \sum_{p=1}^k \frac{(AS'B)_{ip} S_{ip}^2}{S_{ip}^2} \geq \text{Tr}(S^T A S B)$  holds. It follows by expanding the trace and applying the scalar inequality  $\frac{yx^2}{y} \geq 2x\sqrt{\frac{y}{x}} - y'$ . We now

**Algorithm 1** Contrastive calibration on consensus and complementary multi-view representations ( $C^4MV$ ).

**Input:** multi-view dataset  $\mathcal{X} = \{X_1, X_2, \dots, X_{n_v}\}$ , dimension of latent space  $r$ , diversity parameter  $\lambda_1$ , attractive parameter  $\lambda_2$ , repulsive parameter  $\lambda_3$ , fusion parameter  $\alpha$ , and number of maximum iteration  $T$ .

**Output:** Basis set  $\mathcal{W} = \{W_1, \dots, W_{n_v}\}$ , representation set  $\mathcal{H} = \{H_1, \dots, H_{n_v}\}$ , consensus representation  $H_*$ , complementary representation  $H_{\text{com}}$ , and final representation  $H_{\text{final}}$ .

- 1: Initializing  $\mathcal{W}$ ,  $\mathcal{H}$ , and  $H_*$  randomly;
- 2: Construct contrastive graphs  $C^+$  and  $C^-$  according to (8) and (9);
- 3:  $t = 0$
- 4: **while** not converged **and**  $t < T$  **do**
- 5:   **for** each view **do**
- 6:     Updating basis matrix  $W_v$  according to (19);
- 7:     Updating representation matrix  $H_v$  according to (22);
- 8:   **end for**
- 9:   Updating consensus representation matrix  $H_*$  according to (25);
- 10:    $t = t + 1$ ;
- 11: **end while**
- 12: Calculate multi-view complementary representation  $H_{\text{com}}$  by fusing view representations as  $H_{\text{com}} = \sum_{v=1}^{n_v} H_v / n_v$ ;
- 13: Calculate final representation  $H_{\text{final}}$  by concatenating consensus and complementary representations as  $H_{\text{final}} = [\alpha H_*; (1 - \alpha) H_{\text{com}}]$ .

isolate the part of objective (13) that depends on  $H_v$  while keeping all other variables fixed.

**Theorem 1.** Let

$$\begin{aligned} L(H_v) = & \text{Tr}(-2X_v H_v^T W_v^T + W_v H_v H_v^T W_v^T) + \text{Tr}(H_v H_v^T - 2H_v X_v^T W_v) \\ & + \lambda_1 \text{Tr}(H_v^T H_v / n_v) + \sum_u \text{Tr}(H_u (\lambda_2 L_c^+ + \lambda_3 C^-) H_v^T), \end{aligned} \quad (26)$$

and let  $H'_v$  be the value at the previous iteration, and  $M = \lambda_2 L_c^+ + \lambda_3 C^-$ . Then the function

$$\begin{aligned} Z(H_v, H'_v) = & \sum_{i,j} \left[ - (W_v^T X_v)_{ij} (H'_v)_{ij} \left( 1 + \log \frac{(H_v)_{ij}}{(H'_v)_{ij}} \right) + \frac{(H'_v (W_v^T W_v))_{ij} (H'_v)_{ij}^2}{(H'_v)_{ij}} \right. \\ & + \frac{(H'_v)_{ij} (H_v)_{ij}^2}{(H'_v)_{ij}} - (X_v^T W_v)_{ij} (H'_v)_{ij} \left( 1 + \log \frac{(H_v)_{ij}}{(H'_v)_{ij}} \right) \\ & - \sum_{u=1}^{n_v} (H_u M^-)_{ij} (H'_v)_{ij} \left( 1 + \log \frac{(H_v)_{ij}}{(H'_v)_{ij}} \right) + \sum_{u=1}^{n_v} \frac{(H_u M^+)_{ij} (H_v)_{ij}^2}{(H'_v)_{ij}} \\ & \left. + \frac{\lambda_1}{2n_v} (H_v)_{ij} \frac{((H_v)_{ij}^2 + (H'_v)_{ij}^2)}{2(H'_v)_{ij}} \right], \end{aligned} \quad (27)$$

is an auxiliary function for  $L(H_v)$ . Moreover,  $Z$  is convex in  $H_v$  and its minimizer under Nonnegativity constraints are exactly the multiplicative update given in (22).

**Proof.** Each term of  $L(H_v)$  is upper-bounded using Lemma 3 and the inequalities  $z \geq 1 + \log z$  and  $a \leq \frac{a^2 + b^2}{2b}$ . Summing the bounds yields  $Z(H_v, H'_v) \geq L(H_v)$  and equality holds when  $H_v = H'_v$ . Convexity follows from quadratic separability, and minimizing  $Z$  gives the multiplicative update.

**Theorem 2.** Updating  $H_v$  using the rule in (22) monotonically decreases the full objective (13).

**Proof.** Lemma 1 and Theorem 1 imply  $L(H_v^0) = Z(H_v^0, H_v^0) \geq Z(H_v^1, H_v^0) \geq L(H_v^1) \geq \dots$ , so  $L(H_v)$  is monotonically decreasing.

Theorems 1 and 2 together guarantee the convergence of  $C^4MV$  with respect to  $H_v$ . Since the update rules in (19), (22), and (25) satisfy the KKT conditions, the overall objective (13) decreases monotonically and converges.

### 3.6. Complexity analysis

The computational complexity of the proposed  $C^4MV$  model is dominated by the iterative updates of the basis matrices  $W_v$ , the view-specific

**Table 2**  
Summary of the characteristics of the datasets used in this study.

Dataset	#samples	#features in each view	#classes	Application
3-Sources	169	{3560, 3631, 3068}	6	news article classification
BBCSport	544	{3183, 3203}	5	text clustering
Caltech-101	2906	{784, 144, 213}	10	object recognition
Coil100	7200	{30, 30, 30}	100	object recognition
Cora	2708	{1432, 2223}	7	citation network classification
Texas	187	{187, 1703}	5	community detection
Wisconsin	265	{265, 1703}	5	community detection
Washington	230	{230, 1703}	5	community detection
Cornell	183	{183, 1703}	5	community detection

representations  $H_v$ , and the shared consensus representation  $H_s$ . Updating  $W_v$  and  $H_s$  each requires  $\mathcal{O}(dnr)$  operations per iteration, while updating  $H_v$  involves contrastive regularization with dense similarity graphs, resulting in an additional  $\mathcal{O}(rn^2)$  cost due to matrix multiplications with  $C^+$  and  $C^-$ . The total complexity per iteration is therefore  $\mathcal{O}(dnr + rn^2)$ , and over  $T$  iterations, the training complexity becomes  $\mathcal{O}(T(dnr + rn^2))$ . This matches the computational cost of state-of-the-art NMF-based multi-view representation learning methods, demonstrating that C<sup>4</sup>MV achieves its improvements without increasing asymptotic complexity.

#### 4. Experimental results

In this section, we present a comprehensive evaluation of the proposed C<sup>4</sup>MV method through extensive experiments. We assess its performance on nine multi-view datasets using four standard evaluation metrics: Accuracy, NMI, ARI, and F1-score. These metrics respectively evaluate clustering correctness, label correlation, pairwise clustering agreement, and the balance between precision and recall, offering a well-rounded assessment across datasets with diverse characteristics. The results are compared with those of 12 well-established multi-view representation methods to demonstrate the effectiveness of C<sup>4</sup>MV. The final results represent the average and standard deviation over 10 runs, each consisting of 400 iterations of our method. Following unsupervised representation learning, K-means clustering is applied on the representations learned by all methods except CCNMF [25], which uses spectral clustering. The remainder of this section is organized as follows: Section 4.1 describes the benchmark datasets used; Section 4.2 introduces the baseline methods; Section 4.3 presents and analyzes the main experimental results; Section 4.4 explores parameter sensitivity; Section 4.5 conducts an ablation study; and Section 4.6 discusses convergence behavior.

##### 4.1. Datasets

To evaluate the effectiveness of the proposed method, we conducted experiments on nine widely used multi-view datasets spanning various domains, including text, images, and web data. A detailed description of each dataset is provided as follows:

- **3Sources**: Consists of 169 news articles from BBC, Reuters, and The Guardian, each represented in three views according to the source. This dataset is useful for evaluating text-based clustering algorithms focusing on source differentiation and consensus.
- **BBCSport**: This data set includes 544 documents from the BBC Sport website, organized into five topics such as tennis and football. Documents feature two views, text and metadata, which makes them suitable for text clustering and topic analysis.
- **Caltech-101**: A widely-used image dataset consisting of 8677 images categorized into 101 classes. Multi-view representations are typically derived from different feature types such as Gabor features, wavelet moments, and CENTRIST descriptors, supporting visual clustering studies.

- **Coil100**: Contains 7200 images of 100 different objects captured at pose intervals of 5 degrees. Each object is photographed from multiple views, making it a valuable benchmark for evaluating rotational and view variance clustering methods.
- **Cora**: Features scientific papers with each document represented by its textual content and citation links, grouped into thematic classes. This dual-view configuration benefits clustering algorithms that incorporate content and citation data.
- **Texas**: A text dataset where documents are associated with different topics and represented with multiple feature types (e.g., content features and structural features), ideal for testing multi-view representation algorithms that combine text and network information.
- **Wisconsin, Washington, Cornell (WebKB datasets)**: These datasets include web pages from university computer science departments, classified as faculty and students. They are utilized for web page classification, leveraging multi-view data from text, HTML structure, and URL metadata.

Summary statistics and detailed specifications for each dataset are provided in Table 2.

##### 4.2. Compared methods

To assess the performance of the proposed method, the C<sup>4</sup>MV is benchmarked against various leading multi-view representation techniques. A brief overview of these methods is provided below:

- **DiNMF** [22] integrates a diversity term to enhance the distinctiveness of features extracted from different views.
- **LP-DiNMF** [22] extends DiNMF by incorporating local geometric structures of each view, ensuring that the local neighborhood information is preserved.
- **2CMV** [23] learns consensus and complementary information by employing CMF, optimizing for a joint representation.
- **NMF-CC** [40] employs co-orthogonal constraints to enforce orthogonal relationships between features of each view, aiming to reduce redundancy.
- **RRNMF-HSIC** [6] introduces a non-redundancy regularization based on manifold learning, using HSIC to preserve data manifold structures across views.
- **CCNMF** [25] focuses on extracting both consensus and unique features across views using regularization terms customized for each view's contribution.
- **RDINMF** [43] enhances robustness to noise by combining adaptive graph learning with diversity regularization across views.
- **ECNMF** [26] leverages exclusivity and consistency principles to separate and integrate multi-view data features.
- **PRDNMF** [15] utilizes a robust dual graph approach, emphasizing pairwise relationships and co-regularization to preserve the manifold structures of both the data and feature spaces.
- **ADGNMF** [19] unifies autoencoder-like NMF with dual-graph constraints on data and feature manifolds to learn stronger low-dimensional representations for multi-view clustering.

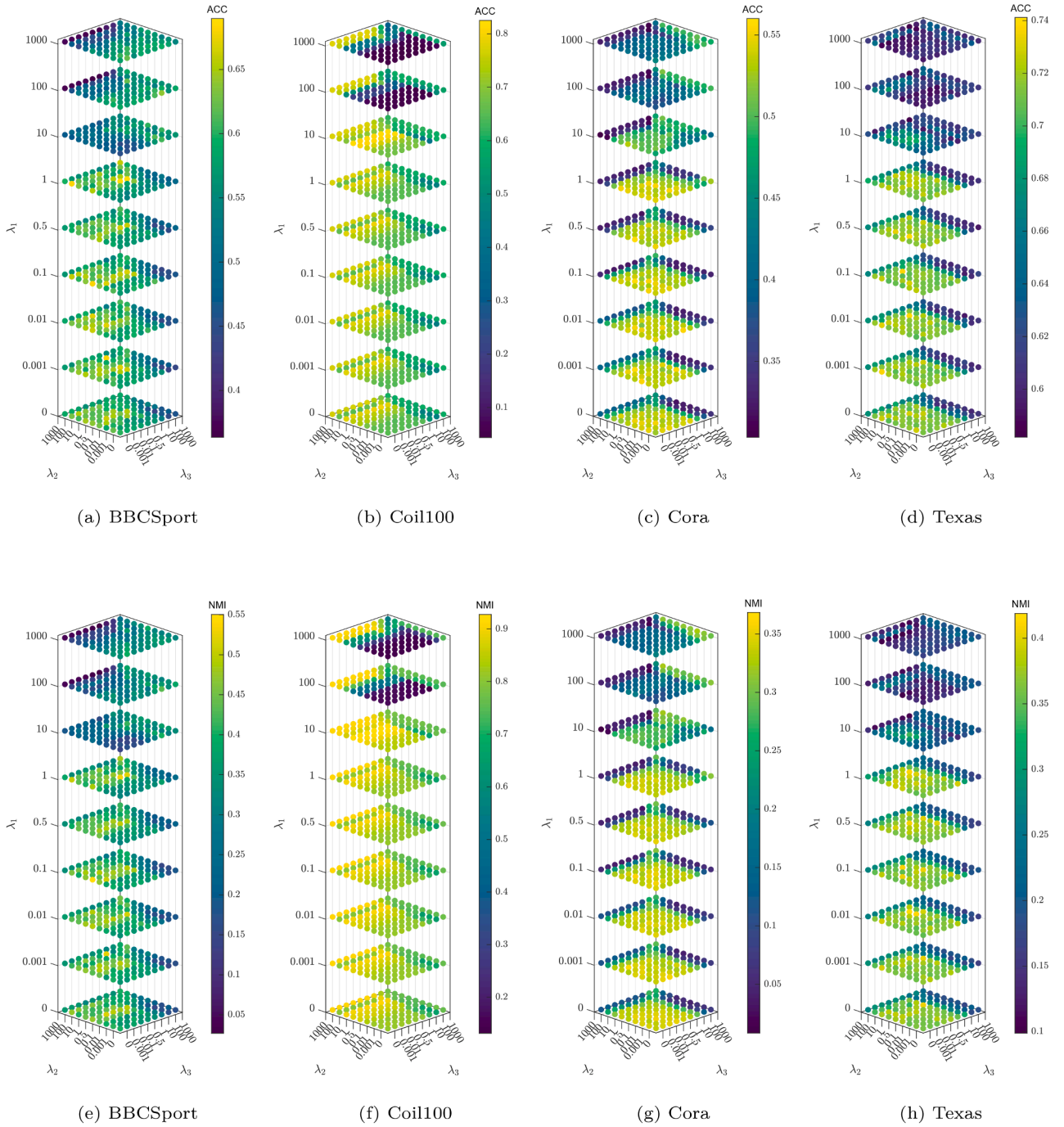
**Table 3**  
Clustering performance comparison across nine multi-view datasets using various algorithms. The top-performing and runner-up results are highlighted in **bold** and underlined, respectively.

Methods	3Sources	BBCSport	Caltech101	Coil100	Cora	Texas	Wisconsin	Washington	Cornell
<b>NMI</b>									
DiNMF	0.2424	0.2195	0.5643	0.8196	0.2820	0.2356	0.2895	0.3009	0.2254
LP-DiNMF	0.2554	0.2765	0.5219	0.8111	0.2245	0.2625	0.2691	0.2559	0.2059
2CMV	<u>0.5147</u>	<u>0.5121</u>	0.5816	0.8809	0.2257	0.2869	0.2392	0.27659	0.2026
NMF-CC	0.3581	0.3485	0.5684	0.7387	0.3035	<u>0.4108</u>	0.4047	<u>0.4107</u>	0.3398
RRNMF-HSIC	0.3613	0.4042	0.5577	0.8095	0.1712	0.2900	0.3741	0.4000	0.3178
CCNMF	0.3931	0.3612	<u>0.5920</u>	0.8456	0.2419	0.3168	0.2976	0.3669	0.2010
RDinNMF	0.2906	0.2781	0.5061	0.7834	0.2366	0.2962	0.3445	0.3408	0.2998
ECNMF	0.2262	0.0802	0.5912	<u>0.9212</u>	0.0283	0.1892	0.1605	0.2368	0.1786
PRDNMF	0.1197	0.3164	0.1466	0.9073	0.0134	0.1243	0.1466	0.3164	0.1197
ADGNMF	0.4056	0.3225	0.5898	0.8938	<u>0.3056</u>	0.3463	<u>0.4226</u>	0.3432	<u>0.3659</u>
ODNMF	0.1667	0.2585	0.5674	0.8191	0.2674	0.3269	0.3473	0.3164	0.2831
SparseMVC	0.4351	0.4340	0.5604	0.8704	0.2809	0.3416	0.4186	0.3595	0.2165
<b>C<sup>4</sup>MV</b>	<b>0.5394</b>	<b>0.5502</b>	<b>0.6123</b>	<b>0.9281</b>	<b>0.3762</b>	<b>0.4243</b>	<b>0.4411</b>	<b>0.4257</b>	<b>0.4236</b>
<b>ACC</b>									
DiNMF	0.4757	0.4922	0.7035	0.6561	0.5038	0.6598	0.6596	0.6817	0.5374
LP-DiNMF	0.5159	0.5216	0.7087	0.6318	0.4703	0.6491	0.6249	0.6391	0.5497
2CMV	<u>0.6721</u>	0.6694	0.7116	0.7418	0.4536	0.6727	0.6203	0.6800	0.5323
NMF-CC	0.5786	0.6025	0.7059	0.5724	0.5548	<u>0.7133</u>	<u>0.7403</u>	0.7191	0.6635
RRNMF-HSIC	0.5893	0.5985	0.7000	0.6499	0.3929	0.6866	0.6928	<u>0.7208</u>	0.6276
CCNMF	0.5763	0.5709	0.7146	0.6896	0.4510	0.7048	0.6716	0.7139	0.5723
RDinNMF	0.5431	0.5426	0.6785	0.6261	0.4372	0.6812	0.6769	0.7060	0.6194
ECNMF	0.5325	0.4238	0.7061	<u>0.7909</u>	0.3070	0.6385	0.5652	0.6565	0.5302
PRDNMF	0.4781	0.3775	0.4381	0.7881	0.3079	0.6042	0.6060	0.6982	0.4902
ADGNMF	0.6142	0.5860	<u>0.7147</u>	0.7724	0.5367	0.6930	0.7260	0.6973	<u>0.6666</u>
ODNMF	0.4520	0.5231	<u>0.6957</u>	0.6175	0.4826	0.6930	0.6701	0.6713	0.5969
SparseMVC	0.5858	<u>0.6820</u>	0.4174	0.6963	<b>0.6043</b>	0.6471	0.6906	0.5696	0.5282
<b>C<sup>4</sup>MV</b>	<b>0.7041</b>	<b>0.7084</b>	<b>0.7276</b>	<b>0.8249</b>	<u>0.5926</u>	<b>0.7465</b>	<b>0.7539</b>	<b>0.7443</b>	<b>0.6953</b>
<b>ARI</b>									
DiNMF	0.1040	0.0946	0.3757	0.5358	0.1433	0.2454	0.2623	0.2354	0.1496
LP-DiNMF	0.1481	0.1427	0.3735	0.5245	0.1350	0.2383	0.2791	0.2031	0.1256
2CMV	0.3218	0.3720	0.4016	0.6434	0.0759	0.2877	0.2243	0.3206	0.1159
NMF-CC	0.2124	0.2284	0.3649	0.4435	0.2335	0.3220	<u>0.3892</u>	0.3786	0.2646
RRNMF-HSIC	0.1816	0.2348	0.3831	0.5427	0.0756	0.2014	<u>0.2972</u>	<u>0.4281</u>	0.3124
CCNMF	0.2342	0.2178	0.4095	0.5750	0.0841	0.2632	0.2668	0.3807	0.2017
RDinNMF	0.1761	0.1595	0.3156	0.5158	0.0699	0.2334	0.2757	0.3681	0.2611
ECNMF	0.1352	0.0520	<u>0.4658</u>	0.6872	0.0141	0.1697	0.0795	0.2867	0.1303
PRDNMF	0.0562	0.0118	0.0861	0.6675	0.0024	0.1537	0.1367	0.3845	0.0845
ADGNMF	0.2127	0.2090	0.4035	<u>0.6963</u>	0.2339	<u>0.3593</u>	0.3629	0.3427	<u>0.3613</u>
ODNMF	0.1016	0.1655	0.3683	0.4985	0.1216	0.2989	0.2693	0.2761	0.2005
SparseMVC	<u>0.3255</u>	<u>0.4026</u>	0.3768	0.6498	<b>0.2685</b>	0.3120	0.3703	0.3024	0.1914
<b>C<sup>4</sup>MV</b>	<b>0.4613</b>	<b>0.4211</b>	<b>0.5411</b>	<b>0.7608</b>	<u>0.2440</u>	<b>0.4333</b>	<b>0.4373</b>	<b>0.4761</b>	<b>0.4414</b>
<b>F1</b>									
DiNMF	0.3873	0.4155	0.6528	0.6013	0.4547	0.5683	0.5943	0.6046	0.4340
LP-DiNMF	0.4319	0.4317	0.6650	0.5723	0.4002	0.5462	0.5394	0.5544	0.4536
2CMV	0.6107	0.6149	0.6587	0.6993	0.3963	0.6111	0.5358	0.6130	0.4671
NMF-CC	0.4827	0.5565	<u>0.6702</u>	0.5280	0.5290	<u>0.6838</u>	<u>0.6978</u>	<u>0.6591</u>	0.5930
RRNMF-HSIC	0.5126	0.5261	0.6326	0.6044	0.3164	0.6178	0.6443	0.6530	0.5458
CCNMF	0.4971	0.4965	0.6697	0.6352	0.3760	0.6386	0.5999	0.6451	0.4873
RDinNMF	0.4574	0.4748	0.6227	0.5775	0.3631	0.5984	0.6023	0.6295	0.5496
ECNMF	0.4347	0.3406	0.6612	0.7404	0.1563	0.5483	0.4874	0.5970	0.4272
PRDNMF	0.3735	0.2763	0.3298	<u>0.7520</u>	0.1537	0.5074	0.5165	0.6275	0.3942
ADGNMF	0.5124	0.5232	0.6642	0.7307	0.4844	0.6455	0.6805	0.6280	<u>0.5958</u>
ODNMF	0.3351	0.4420	0.6546	0.5570	0.4029	0.61141	0.6059	0.6011	0.5151
SparseMVC	<u>0.6259</u>	<b>0.6823</b>	0.6455	0.6701	<b>0.5804</b>	0.6322	0.6798	0.6504	0.4739
<b>C<sup>4</sup>MV</b>	<b>0.6662</b>	<u>0.6704</u>	<b>0.6790</b>	<b>0.7898</b>	<u>0.5636</u>	<b>0.6956</b>	<b>0.7092</b>	<b>0.6960</b>	<b>0.6339</b>

- ODNMF [44] enforces cross-view and basis-matrix orthogonality to preserve viewpoint characteristics and reduce redundancy, while graph regularization captures intrinsic structure.
- SparseMVC [50] adaptively adjusts view encodings with entropy-matched sparsity, reweights samples via attention-based correlation, and aligns cross-view feature distributions for flexible multi-view clustering.

### 4.3. Results

This section evaluates the representation capability of the proposed C<sup>4</sup>MV model by comparing its clustering results with 12 state-of-the-art multi-view representation methods across nine real-world datasets. Table 3 reports the outcomes, where the best performance for each metric is highlighted in **bold** and the second best is underlined. The

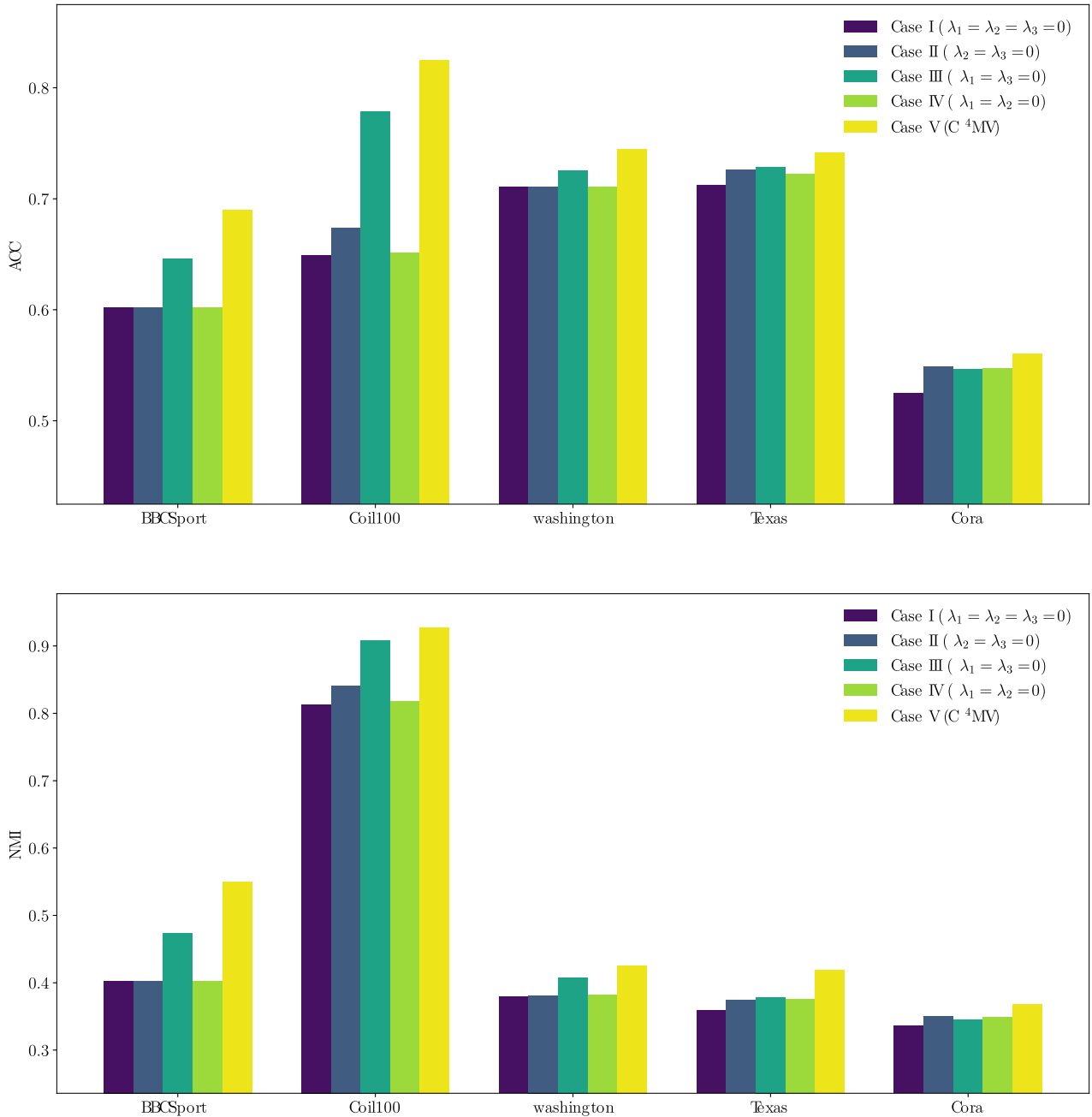


**Fig. 3.** Parameter analysis of  $C^4MV$  with respect to  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  on four datasets. Subfigures (a)–(d) show ACC results; (e)–(h) show NMI. Lighter colors indicate higher values.

results reveal that  $C^4MV$  consistently learns effective multi-view representations that enhance clustering performance. From Table 3, we observe:

- The proposed  $C^4MV$  method consistently achieves superior clustering performance, obtaining the highest performance across most datasets. Specifically, it obtains the best results in 32 out of 36 evaluation cases across all datasets. On average,  $C^4MV$  surpasses the second-best method by approximately 3.2%, 2.14%, 5.8%, and 1.74% in the NMI, ACC, ARI, and F1 scores, respectively, confirming its effectiveness in extracting discriminative and consensus-preserving representations.

- On datasets characterized by moderate redundancy and complementary structure, such as *3Sources* and *BBCSport*,  $C^4MV$  significantly enhances performance across all metrics. For example, on *3Sources*, ACC improves from 0.672 (2CMV) to 0.704 and ARI rises from 0.325 (SparseMVC) to 0.461, while on *BBCSport*, NMI reaches 0.550 and ACC climbs to 0.708, representing absolute gains of approximately 2.6 to 4% in most metrics, with a markedly larger 13.6% improvement in ARI on *3Sources*. These results show that separating and then combining shared and unique information using contrastive calibration clearly improves performance.
- The same representational mechanism extends effectively to visual datasets. On *Caltech101* and *Coil100*, the  $C^4MV$  model consistently



**Fig. 4.** Ablation study (in terms of NMI and ACC measures) of the  $C^4MV$  algorithm with respect to the parameters  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  on five real-world datasets. Case I ( $\lambda_1 = \lambda_2 = \lambda_3 = 0$ ) corresponds to the baseline model with all regularizations disabled; Case II ( $\lambda_2 = \lambda_3 = 0$ ) examines the effect of diversity term alone; Case III ( $\lambda_1 = \lambda_3 = 0$ ) examines the effect of attractive term alone; Case IV ( $\lambda_1 = \lambda_2 = 0$ ) examines the effect of repulsive term alone; and Case V ( $C^4MV$ ) corresponds to the full model with all regularizations active.

achieves the highest performance across all metrics. Specifically, on *Coil100*,  $C^4MV$  surpasses the second-best method by considerable margins, including a notable absolute improvement of approximately 6.5% in ARI. Additionally, the model demonstrates robust, balanced performance improvements in all metrics compared to previous methods, underscoring the effectiveness of integrating contrastive attraction and repulsion mechanisms, particularly beneficial for visual datasets containing many visually similar classes.

- Moreover, on graph-based datasets like *Cora* and subsets from We-bKB (*Texas*, *Wisconsin*, *Washington*, and *Cornell*),  $C^4MV$  again attains top performance. Specifically, the ARI significantly improves on *Texas*, 0.359 (ADGNMF) to 0.433 ( $C^4MV$ ), representing an absolute gain of

approximately 7.4%. On *Wisconsin*, ARI rises from 0.389 (NMF-CC) to 0.437, marking an absolute improvement of around 4.8%. These substantial improvements indicate that the  $C^4MV$  model effectively manages high sparsity and noise in datasets, where the performance of competing methods can vary considerably.

- Overall, the results across nine diverse datasets consistently validate the representational design of  $C^4MV$ . By explicitly separating consensus and complementary information, the model reduces interference between shared and view-specific cues, while contrastive calibration sharpens cluster boundaries and strengthens discriminative structure. This unified mechanism produces cleaner embeddings and more stable clustering across modalities,

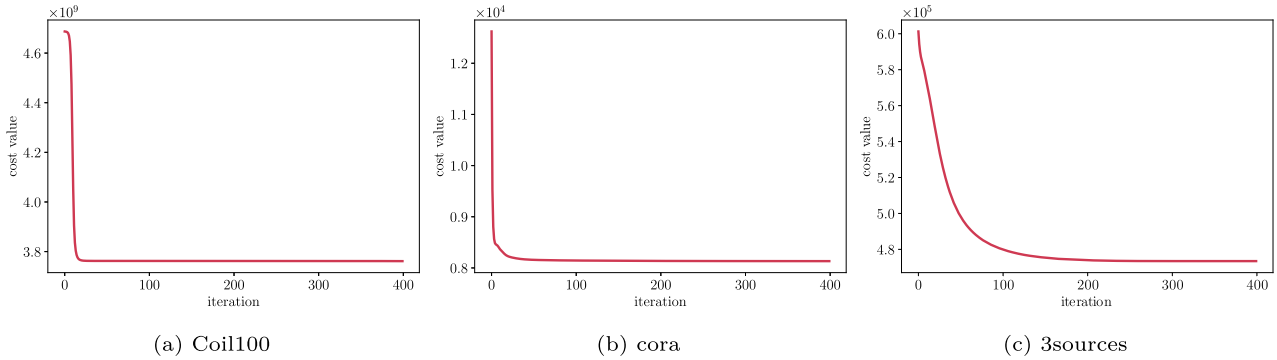


Fig. 5. Convergence analysis of the  $C^4MV$  algorithm on three datasets.

accounting for the strong and consistent performance reported in Table 3.

#### 4.4. Parameter analysis

In this subsection, we will evaluate the effects of the hyperparameters in the proposed algorithm. To evaluate the sensitivity and effectiveness of the proposed  $C^4MV$  framework, we conduct a comprehensive analysis on its three core hyperparameters: the diversity regularization hyperparameter  $\lambda_1$ , the attractive contrastive regularization hyperparameter  $\lambda_2$ , and the repulsive contrastive regularization hyperparameter  $\lambda_3$ . These hyperparameters govern the trade-off between consensus and complementary representation learning, as well as the strength of the contrastive calibration mechanism. Fig. 3 visualizes the variation in cluster performance, measured by ACC and NMI, in four data sets. The hyperparameters are selected from a broad value range  $\{0, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1, 10^1, 10^2, 10^3\}$  to capture the full behavior spectrum of the model under different regularization settings. In the Fig. 3, yellow indicates better performance (higher ACC/NMI) while blue indicates worse performance (lower ACC/NMI). The experimental findings can be summarized as follows:

- **Diversity regularizer  $\lambda_1$ :** Moderate values of  $\lambda_1$  (e.g.,  $\{0.1, 0.5, 1\}$ ) yield optimal performance by encouraging orthogonality between consensus and complementary components. This trend holds across both graph-structured datasets (e.g., *Cora*, *Texas*) and text corpora (e.g., *BBCSport*). In contrast, excessively large values (e.g.,  $\{100, 1000\}$ ) overly suppress shared structure, leading to performance degradation. Moreover, since  $\lambda_1$  and  $\lambda_2$  exert opposing effects, smaller values tend to offer more stable and insensitive behavior, especially when both are active.
- **Attractive regularizer  $\lambda_2$ :** Increasing  $\lambda_2$  consistently enhances alignment among semantically related samples across views. This results in steady performance gains, particularly on graph datasets like *Cora* and *Texas*. The attractive force becomes more effective when more positive sample pairs are available, strengthening cross-view consistency. However, excessively large  $\lambda_2$  may interfere with high  $\lambda_1$ , reducing overall effectiveness.
- **Repulsive regularizer  $\lambda_3$ :** The impact of  $\lambda_3$  varies by data type. On visual datasets such as *Coil100*, moderate repulsion (e.g.,  $\{1, 10\}$ ) effectively prevents representation collapse by pushing apart dissimilar instances, leading to performance gains. For structured (graph or text) datasets, however, the repulsive term contributes only marginal improvements.

For the *Coil100* dataset, the model achieves its highest ACC of 0.80 (yellow in the figure) when  $\lambda_1 = 10$ ,  $\lambda_2 = 10$ , and  $\lambda_3 = 0.001$ , indicating optimal performance. Decreasing  $\lambda_2$  to 1 or 0.5 reduces ACC to 0.75 and 0.70 (greenish), while a very small  $\lambda_2 = 0.1$  leads to poor performance of 0.55 (bluish). These results highlight that balanced weights on con-

sensus and complementary representations, along with appropriate contrastive calibration, are crucial for achieving high clustering accuracy.

To balance the contribution of consensus and complementary information in the final representation, we introduce a weighting parameter  $\alpha$  that controls the trade-off between shared and view-specific features. We evaluate  $\alpha$  over the range  $[0, 1]$  using discrete values  $\{0, 0.25, 0.5, 0.75, 1\}$ . Experimental results indicate that intermediate values (e.g., 0.5, 0.75) consistently achieve superior performance, underscoring the importance of integrating both consensus and complementary cues. In contrast, relying exclusively on either component (0 or 1) leads to reduced performance across various datasets, with stronger effects observed on content-diverse corpora such as *BBCSport*.

Overall, the best results are obtained with  $\lambda_1 \in [0.1, 1]$  and  $\lambda_2, \lambda_3 \in [1, 10]$ , indicating that  $C^4MV$  is robust to precise hyper-parameter settings yet benefits from balanced regularisation. The consistent improvements across diverse datasets validate the effectiveness of jointly leveraging contrastive calibration with consensus and complementary representations learning.

#### 4.5. Ablation study

To evaluate the contribution of individual components within the proposed  $C^4MV$  framework, we conduct comprehensive ablation experiments across five datasets, as illustrated in Fig. 4, using ACC and NMI as evaluation metrics. The analysis primarily focuses on evaluating the impact of consensus and complementary diversity and the contrastive calibration mechanism. We conduct an ablation study involving five progressively enhanced variants of the base model to isolate the contribution of each component. In the most basic version (Case I), where both the diversity and contrastive calibration terms are removed, performance deteriorates substantially, underscoring the importance of disentangling the shared and view-specific information. Introducing only the consensus and complementary diversity (Case II), without contrastive terms, already yields significant performance gains, highlighting the structural advantage of explicitly modeling this distinction. When either contrastive attraction (Case III) or repulsion (Case IV) is introduced in isolation, additional performance improvements are observed. Specifically, contrastive attraction proves more effective for graph-structured datasets such as *Cora*, *Washington*, and *Texas*, while contrastive repulsion yields greater benefits on visual datasets like *Coil100*. In contrast, text-based datasets such as *BBCSport* exhibit only marginal gains from repulsion, likely due to their inherently sparse or semantically entangled structure. Notably, the full  $C^4MV$  model (Case V), which incorporates both contrastive terms alongside consensus-complementary diversity, consistently achieves the highest performance across all datasets. Relative to the baseline (Case I), Case II yields modest gains (1.5% NMI, 3% ACC), while Case III and Case IV achieve larger average improvements of about 4.5%/8.5% and 3%/6.5% in NMI/ACC, respectively. The full model (Case V) delivers the largest gains, improving NMI by

approximately 25% and ACC by about 15%, confirming the complementary effects of structural disentanglement and contrastive calibration.

#### 4.6. Convergence analysis

To empirically assess the convergence behavior of the proposed algorithm, we record the objective function value across 400 iterations for three representative datasets: *Coil100*, *Cora*, and *3Sources*. As illustrated in Fig. 5, the objective value exhibits a consistently decreasing trajectory throughout the optimization process. A significant reduction is observed in the early iterations, followed by a gradual stabilization, indicating convergence toward a stationary point. In all cases, the algorithm reaches a stable solution within the first 100 iterations, suggesting computational efficiency and robustness. The x-axis of the plot represents the number of iterations, while the y-axis corresponds to the value of the objective function. A sharp initial descent in the curve highlights the algorithm's rapid progress during early optimization stages. The consistent convergence trends observed across datasets of different modalities further support the general applicability and reliability of the proposed method in multi-view representation scenarios. These results collectively demonstrate that the  $C^4MV$  optimization procedure is both stable and efficient, with low risk of divergence or oscillation.

## 5. Conclusion

This work proposed a novel framework for unsupervised multi-view representation learning that leverages model-level and late-fusion strategies to effectively capture both consensus and complementary information across views. By employing a combination of joint and disjoint encoder-decoder NMFs along with diversity regularization, this model can extract meaningful shared and view-specific features. A central innovation of our approach is the introduction of contrastive calibration regularization, which improves sample-level discriminability by leveraging cross-view positive pairs through graph-based contrastive learning. This reduces reliance on negative pairs and enhances the alignment of semantically similar samples across views. Furthermore,  $C^4MV$  integrates these model-level and late fusion strategies within a unified, efficient optimization framework. We conducted extensive evaluations on 12 state-of-the-art methods across nine diverse real-world datasets, demonstrating that our method consistently achieves superior performance in the unsupervised representation learning task. These results highlight the effectiveness of our approach across various domains and data types.

While the proposed method shows strong performance, it still requires careful tuning of multiple hyperparameters, which can limit its adaptability. Future work could explore adaptive regularization techniques that adjust the influence of contrastive terms based on model dynamics. Another important direction is to extend  $C^4MV$  into a semi-supervised contrastive learning setting, where limited label information guides the contrastive calibration process. To improve scalability, anchor graph-based methods could be incorporated to approximate neighborhood structures more efficiently in large datasets. Finally, future research may extend  $C^4MV$  toward transform learning-inspired multi-view architectures or contrastive dictionary fusion, providing a natural connection to broader representation learning paradigms and enabling more flexible cross-view representation learning.

#### CRedit authorship contribution statement

**Negin Jabari:** Writing – original draft, Visualization, Software, Methodology, Investigation; **Amjad Seyedi:** Writing – review & editing, Software, Methodology, Investigation, Conceptualization; **Reza Mahmoodi:** Writing – review & editing, Visualization, Validation, Methodology, Data curation; **Fardin akhlaghian Tab:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

#### Data availability

Data will be made available on request.

#### Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Amjad Seyedi reports financial support was provided by the European Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

Amjad Seyedi acknowledges the support by the European Union (ERC consolidator, eLinoR, no 101085607).

#### References

- [1] M.-S. Chen, J.-Q. Lin, X.-L. Li, B.-Y. Liu, C.-D. Wang, D. Huang, J.-H. Lai, Representation learning in multi-view clustering: a literature review, *Data Sci. Eng.* 7 (3) (2022) 225–241.
- [2] J. Chen, G. Wang, G.B. Giannakis, Graph multiview canonical correlation analysis, *IEEE Trans. Signal Process.* 67 (11) (2019) 2826–2838.
- [3] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, D. Xu, Generalized latent multi-view subspace clustering, *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (1) (2018) 86–99.
- [4] Z. Jiao, H. Zhang, X. Li, Deep graph multi-view representation learning with self-augmented view fusion, *IEEE Trans. Neural Netw. Learn. Syst.* 36 (8) (2025) 14119–14130.
- [5] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: *Proceedings of the 2013 SIAM International Conference on Data Mining*, SIAM, 2013, pp. 252–260.
- [6] G. Cui, Y. Li, Nonredundancy regularization based nonnegative matrix factorization with manifold learning for multiview data representation, *Inf. Fusion* 82 (2022) 86–98.
- [7] Z. Zhao, T. Wang, H. Xin, R. Wang, F. Nie, Multi-view clustering via high-order bipartite graph fusion, *Inf. Fusion* 113 (2025) 102630.
- [8] X. Yang, H. Che, M.-F. Leung, Tensor-based unsupervised feature selection for error-robust handling of unbalanced incomplete multi-view data, *Inf. Fusion* 114 (2025) 102693.
- [9] Z. Chen, X.-J. Wu, T. Xu, H. Li, J. Kittler, Multi-layer multi-level comprehensive learning for deep multi-view clustering, *Inf. Fusion* 116 (2025) 102785.
- [10] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* 401 (6755) (1999) 788–791.
- [11] A.P. Singh, G.J. Gordon, Relational learning via collective matrix factorization, *KDD '08*, Association for Computing Machinery, New York, NY, USA, 2008, p. 650–658.
- [12] M. Liu, Z. Yang, L. Li, Z. Li, S. Xie, Auto-weighted collective matrix factorization with graph dual regularization for multi-view clustering, *Knowl. Based Syst.* 260 (2023) 110145.
- [13] S. Shi, F. Nie, R. Wang, X. Li, Multi-view clustering via nonnegative and orthogonal graph reconstruction, *IEEE Trans. Neural Netw. Learn. Syst.* 34 (1) (2023) 201–214.
- [14] L. Zong, X. Zhang, L. Zhao, H. Yu, Q. Zhao, Multi-view clustering via multi-manifold regularized non-negative matrix factorization, *Neural Netw.* 88 (2017) 74–89.
- [15] H. Tang, S. Liu, Y. Tang, F. Yu, Multi-view clustering based on pairwise co-regularization and robust dual graph non-negative matrix factorization, *Neurocomputing* 611 (2025) 128594.
- [16] S. Huang, I. Tsang, Z. Xu, J. Lv, Q.-H. Liu, Multi-view clustering on topological manifold, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 2022, pp. 6944–6951.
- [17] M.M. Kalayeh, H. Idrees, M. Shah, NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 184–191.
- [18] G. Cui, R. Wang, D. Wu, Y. Li, Semi-supervised multi-view clustering based on NMF with fusion regularization, *ACM Trans. Knowl. Discov. Data* 18 (6) (2024) 1–26.
- [19] Y. Ban, Y. Cai, Z. Huang, Autoencoder-like non-negative matrix factorization with dual-graph constraints for multi-view clustering, *Int. J. Mach. Learn. Cybern.* 16 (2025) 5637–5652.
- [20] L. Zhao, Z. Wang, Z. Wang, Z. Chen, Multi-view graph regularized deep autoencoder-like NMF framework, in: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [21] H. Huang, G. Zhou, Q. Zhao, L. He, S. Xie, Comprehensive multiview representation learning via deep autoencoder-like nonnegative matrix factorization, *IEEE Trans. Neural Netw. Learn. Syst.* 35 (5) (2024) 5953–5967.
- [22] J. Wang, F. Tian, H. Yu, C.H. Liu, K. Zhan, X. Wang, Diverse non-negative matrix factorization for multiview data representation, *IEEE Trans. Cybern.* 48 (9) (2017) 2620–2632.
- [23] K. Luong, R. Nayak, A novel approach to learning consensus and complementary information for multi-view data clustering, in: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, IEEE, 2020, pp. 865–876.

- [24] M. Horie, H. Kasai, Consistency-aware and inconsistency-aware graph-based multi-view clustering, in: 2020 28th European Signal Processing Conference (EUSIPCO), IEEE, 2021, pp. 1472–1476.
- [25] G. Li, D. Song, W. Bai, K. Han, R. Tharmarasa, Consensus and complementary regularized non-negative matrix factorization for multi-view image clustering, *Inf. Sci.* 623 (2023) 524–538.
- [26] H. Huang, G. Zhou, Y. Zheng, Z. Yang, Q. Zhao, Exclusivity and consistency induced NMF for multi-view representation learning, *Knowl. Based Syst.* 281 (2023) 111020.
- [27] S.-J. Xiang, H.-C. Li, J.-H. Yang, X.-R. Feng, Dual auto-weighted multi-view clustering via autoencoder-like nonnegative matrix factorization, *Inf. Sci.* 667 (2024) 120458.
- [28] C. Zhang, Y. Geng, Z. Han, Y. Liu, H. Fu, Q. Hu, Autoencoder in autoencoder networks, *IEEE Trans. Neural Netw. Learn. Syst.* 35 (2) (2022) 2263–2275.
- [29] B.-J. Sun, H. Shen, J. Gao, W. Ouyang, X. Cheng, A non-negative symmetric encoder-decoder approach for community detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 597–606.
- [30] J. Maggu, A. Majumdar, Semi-coupled transform learning, in: L. Cheng, A.C.S. Leung, S. Ozawa (Eds.), *Neural Information Processing*, Springer International Publishing, Cham, 2018, pp. 141–150.
- [31] J. Maggu, A. Goel, R. Kumar, Label-consistent kernel transform learning-based sparse hashing for cross-modal retrieval, *Knowl. Inf. Syst.* 67 (2025) 6937–6967.
- [32] G.E. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Comput.* 14 (8) (2002) 1771–1800.
- [33] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2019. 1807.03748
- [34] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
- [35] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 1597–1607.
- [36] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, in: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, Springer, 2020, pp. 776–794.
- [37] K. Hassani, A.H. Khasahmadi, Contrastive multi-view representation learning on graphs, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 4116–4126.
- [38] X. Yang, J. Jiaqi, S. Wang, K. Liang, Y. Liu, Y. Wen, S. Liu, S. Zhou, X. Liu, E. Zhu, DealMVC: dual contrastive calibration for multi-view clustering, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 337–346.
- [39] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, *Inf. Fusion* 91 (2023) 424–444.
- [40] N. Liang, Z. Yang, Z. Li, W. Sun, S. Xie, Multi-view clustering by non-negative matrix factorization with co-orthogonal constraints, *Knowl. Based Syst.* 194 (2020) 105582.
- [41] L. Feng, W. Liu, X. Meng, Y. Zhang, Re-weighted multi-view clustering via triplex regularized non-negative matrix factorization, *Neurocomputing* 464 (2021) 352–363.
- [42] G.A. Khan, J. Hu, T. Li, B. Diallo, H. Wang, Multi-view data clustering via non-negative matrix factorization with manifold regularization, *Int. J. Mach. Learn. Cybern.* 13 (3) (2022) 677–689.
- [43] C. Li, H. Che, M.-F. Leung, C. Liu, Z. Yan, Robust multi-view non-negative matrix factorization with adaptive graph and diversity constraints, *Inf. Sci.* 634 (2023) 587–607.
- [44] X. Zhang, C. Leng, J. Peng, I. Cheng, A. Basu, Orthogonal diversity nonnegative matrix factorization for multi-view clustering, *Eng. Appl. Artif. Intell.* 152 (2025) 110715.
- [45] H. Yuan, Z. Zhang, Q. Guo, L. Chi, S. Ruan, W. Zhou, J. Pang, X. Hao, DWCL: dual-weighted contrastive learning for robust multi-view clustering, *Eng. Appl. Artif. Intell.* 165 (2026) 113532.
- [46] X. Huang, R. Zhang, Y. Li, F. Yang, Z. Zhu, Z. Zhou, MFC-ACL: multi-view fusion clustering with attentive contrastive learning, *Neural Netw.* 184 (2025) 107055.
- [47] W. Wu, D. Wang, M. Wang, S. Feng, Y. Zhang, Sentiment triplet extraction with multi-view contrastive learning, *IEEE Trans. Affect. Comput.* 16 (2024) 1–18.
- [48] J. Zhu, X. Zou, L. Liu, Z. Huang, Y. Zhang, C. Tang, L.-R. Dai, Trusted Mamba contrastive network for multi-view clustering, in: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.
- [49] Z. Dong, J. Jin, Y. Xiao, B. Xiao, S. Wang, X. Liu, E. Zhu, Subgraph propagation and contrastive calibration for incomplete multiview data clustering, *IEEE Trans. Neural Netw. Learn. Syst.* 36 (2) (2025) 3218–3230.
- [50] R. Liu, X. Zou, C. Tang, X. Zheng, X. Hu, K. Sun, X. Liu, SparseMVC: probing cross-view sparsity variations for multi-view clustering, in: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [51] W. Barkhoda, A. Seyedi, N. Gillis, F. Akhlaghian Tab, Instance-wise distributionally robust nonnegative matrix factorization, *Pattern Recognit.* 169 (2026) 111732.
- [52] S. Ghodsi, S.A. Seyedi, E. Ntoutsis, Towards cohesion-fairness harmony: contrastive regularization in individual fair graph clustering, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2024, pp. 284–296.
- [53] R. Mahmoodi, S.A. Seyedi, A. Abdollahpouri, F. Akhlaghian Tab, Enhancing link prediction through adversarial training in deep nonnegative matrix factorization, *Eng. Appl. Artif. Intell.* 133 (2024) 108641.
- [54] N. Salahian, F.A. Tab, S.A. Seyedi, J. Chavoshinejad, Deep autoencoder-like NMF with contrastive regularization and feature relationship preservation, *Expert Syst. Appl.* 214 (2023) 119051.
- [55] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: *Proceedings of the 14th International Conference on Neural Information Processing Systems*, 13, 2000, pp. 535–541.