Community-aware Multi-view Representation Learning with Incomplete Information

Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, Xi Peng

Abstract—Due to the complexity of data collection in the real world, Multi-view Representation Learning (MvRL) always encounters the *incomplete information* challenge referred to as the Sample-missing Problem (SP) and the View-unaligned Problem (VP). Although several methods have been proposed, they fail to find a good trade-off among sample restoration, view alignment, and data diversity preservation. To address this issue, we take and mathematically formulate two sociological concepts for MvRL, *i.e.*, *community commonality* and *community versatility*, where the former refers to the identical custom shared within the same community, and the latter refers to the similar but non-identical custom within communities of the same minority. One could find that the *community commonality* can enhance the compactness of view-specific clusters, and the *community versatility* can preserve the view diversity. What is more important, both of them together could be helpful to MvRL with the incomplete information. With the formulations, we propose a novel method dubbed Community-Aware Multi-viEw RepresentAtion learning with incomplete information (CAMERA). In brief, CAMERA employs a novel dual-stream network and an elaborate objective function that theoretically and empirically embraces community commonality and versatility. Extensive experimental results on seven datasets demonstrate that CAMERA remarkably outperforms 24 competitive multi-view learning methods on clustering, classification, and human action recognition tasks. The code is available at https://github.com/XLearning-SCU/2025-TPAMI-CAMERA.

Index Terms—Multi-view Learning, Incomplete Information, Community Commonality, Community Versatility.

1 Introduction

Toward achieving robustness against SP and VP, a number of methods have been proposed. To tackle the SP problem, some methods impute the missing samples with the help of neighboring counterparts [1] or generative models [2], [3]. To address the VP challenge, most of the existing approaches resort to establishing the correspondence between cross-view samples through instance identification [4] or graph matching [5]. Despite the promising performance achieved by these methods, most of them can only tackle either SP or VP. To the best of our knowledge, SURE [4] would be one of the few solutions to handle SP and VP under a unified framework. In short, SURE assumes that all samples could be mapped into a common space, wherein the cross-view neighbors are used to impute missing samples and establish correspondence for unaligned views. As SURE utilizes the cross-view sample-sample relationship in a common space to restore incomplete information, it is inevitable to overemphasize the cross-view consistency and lose the data diversity. From the above discussion, one could conclude that it is quite difficult to find a good trade-off among sample restoration, view alignment, and data diversity preservation for the incomplete MvRL methods.

To better tackle the incomplete information challenge faced by MvRL while preserving data diversity and avoiding overemphasizing the cross-view consistency, we bring the *community* concept from sociology [6] and formally formulate it into MvRL. As shown in Fig. 1, in sociology, *community* [7] refers to a group of people gathering in a specific region, and multiple region-specific communities form a minority. Generally speaking, the identical custom [8] shared within the same community (dubbed *community commonality*) benefits the community cohesion, while the similar but non-identical custom of communities of the same minority (dubbed *community versatility*) helps to establish the ethnic correspondence and preserve the minority diversity. Both the commonality and the versatility are vital to the continuity and development of the minority. With the above concepts, we propose a novel incomplete MvRL method (dubbed CAMERA) which encapsulates the above two community characteristics into MvRL, where the community, the minority, and the region correspond to the view-specific cluster, the view-integration cluster and the view, respectively. Accordingly, community commonality refers to a view-specific characteristic desired by sample-level representation learning, and community versatility is expected in learning the cross-view cluster-level representation.

To endow MvRL with the two community characteristics, the first thing is to mathematically represent the community. To this end, the most straightforward approach is representing the community using its center obtained by various clustering methods such as k-means [9]. However, such an approach is with sub-optimal performance as proved in our ablation studies, which fails to embrace the community characteristics due to the following reasons: i) it is hard to guarantee the community commonality in learn-

- *This work was supported in part by NSFC under Grant 624B2099, 62472295, U24B20174; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303, CJ202403; in part by Sichuan Science and Technology Planning Project under Grant 24NS-FTD0130; and in part by Baidu Scholarship.*
- *H. Li, Y. Lin, H. Peng, M. Yang, and X. Peng are with the College of Computer Science, Sichuan University, Chengdu 610065, China. X. Peng is also with the National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, China. E-mail: {haobinli.gm, linyijie.gm, penghu.ml, yangmouxing, pengx.gm}@gmail.com.*
- *Corresponding authors: M. Yang and X. Peng.*

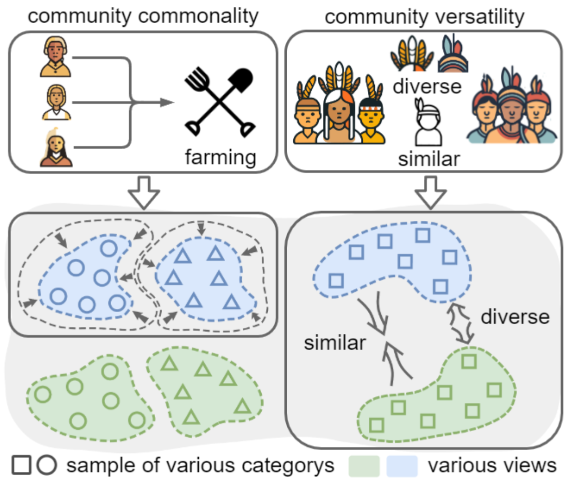


Fig. 1. Our observations. Without loss of generality, we take bi-view data as a showcase to introduce two important concepts in sociology, *i.e.*, community commonality and community versatility. To be specific, community commonality indicates that the people within a community take the common civilization (*e.g.*, farming for a livelihood), while community versatility indicates different communities of the same minority share similar but diverse customs (*e.g.*, wearing hats but with different feathers). With the two sociological concepts, we reveal that two community characteristics could be accordingly derived and a well-established MvRL method always embraces the characteristics, especially when encountering the incomplete information challenge (SP and VP). More specifically, motivated by the common civilization, the within-community samples should share a common pattern, which imposes coherence on the within-community samples. On the other hand, motivated by the community versatility, the communities of the same category should be highly similar instead of identical. Clearly, it is highly expected to find the balance between highly similar and measurably diverse patterns so that the cross-view correspondence could be well established while preserving the complementary information across views. Through embracing the community characteristics, a new paradigm for MvRL with incomplete information could be designed as the major contribution of this paper.

ing sample-level representation since the learned sample-level representation is decoupled with the cluster-level representation. In other words, the cluster-level representation often depends on the sample-level representation, but not vice versa; ii) although it is easy to enforce the cross-view cluster-level representations similar, it keeps unclear how to make them diverse. In other words, it is daunting to guarantee the community versatility during learning cross-view cluster-level representation, so that the view-specific information could be preserved and integrated to facilitate the downstream tasks.

To achieve better incomplete MvRL performance, we design a novel dual-stream network to couple the sample- and cluster-level representation learning processes. In brief, our dual-stream network employs a mutual attention (MA) mechanism to dexterously manipulate the attention between samples and learnable community centers in a dual manner. To be specific, given a sample as the query and community centers as the gallery, our method integrates the retrieved community center with the sample-level representation using the sample-to-cluster attention mechanism, thus enjoying the community commonality. In a coupled fashion, the cluster-level representation is learned from the sample set through the cluster-to-sample attention mechanism. Besides the contribution in the network architecture, we elaborately design a new objective function that endows the cross-view cluster-level representation with community versatility while further enhancing community commonality for sample-level representation. In short, the proposed objective function consists of a community commonality learning loss, a sample consistency learning loss, and a community versatility learning loss. In brief, the commonality learning loss imposes coherence to the view-specific within-cluster samples, thus enhancing the compactness of clusters. The sample consistency learning loss aims to enforce the cross-view samples of the same instance similar, which favors learning view-integration cluster representation. As the other term of our objective, the versatility learning loss is designed to learn the cross-view cluster-level representation in a non-monotonic manner, which could avoid overemphasizing the cross-view consistency and preserve the view diversity.

Thanks to the community commonality and the community versatility, our method finds an elegant balance among sample restoration, view alignment, and data diversity preservation, thus benefiting the incomplete MvRL. In summary, the contributions and novelties of this work could be summarized as follows:

- Motivated by two concepts in sociology, we reveal that community commonality could enhance the compactness of view-specific clusters, and community versatility would preserve the multi-view diversity. The introduction and employment of the two sociological concepts show a feasible way and novel insights toward achieving the robustness against incomplete information for MvRL.
- To implement community commonality and community versatility, we propose a novel dual-stream network with an elaborate objective function to learn view-specific sample-level representation and cross-view cluster-level representation. The theoretical analysis demonstrates that the objective function could capture both the community commonality and versatility from an informative perspective.
- To restore the incomplete information by taking advantage of community characteristics, we propose a novel data imputation and alignment method under a unified MA-based framework. Extensive experiments verify that community commonality and versatility could boost the performance in incomplete information restoration.

2 Related Works

In this section, we briefly review two topics related to this work, *i.e.*, multi-view representation learning and incomplete information restoration.

2.1 Multi-view Representation Learning

Multi-view representation learning (MvRL) aims to learn a common space for multi-view data, wherein the representations are extracted for handling downstream tasks. Based on the paradigms to construct the common space, MvRL methods could be divided into two categories, namely, i) the regularization-based methods [10]–[12], which learn the common space by employing various regularizations, such as correlation maximization [10], [11], consistent Laplacian graph [13], norm constraints [14], and consensus cluster matrix/structure [15], [16]; ii) the contrastive methods [3], [17], which utilize contrastive loss to simultaneously perform representation learning and achieve cross-view consistency in the common space.

Different from the existing works, this paper explicitly uncovers that the community characteristics could benefit MvRL. On the one hand, community commonality could enhance the compactness of view-specific clusters, thus learning discriminative representations, which is the foundation of MvRL. On the other hand, community versatility enriches representation with unique view-specific information, facilitating multi-view diversity. By exploring and exploiting the community characteristics, CAMERA achieves a new SOTA performance in several multi-view learning tasks, including clustering, classification, and action recognition.

2.2 Incomplete Information Restoration

Most MvRL methods implicitly or explicitly rely on the assumption of complete information. However, this assumption could be undermined, leading to the SP and VP. Toward achieving robust MvRL, some methods have been proposed in the past decade which could be grouped into SP-oriented and VP-oriented methods. The SP-oriented MvRL methods aim to impute the missing samples by leveraging the observed samples, which could be divided into two categories: i) neighborhood-based methods [1], [4], which leverage cross-view nearest neighbors to impute missing samples; ii) generative methods, which learn a global mapping across views and utilize the mapping to impute the missing samples with the help of the observed counterpart samples. In contrast, VP-oriented MvRL methods aim to establish cross-view correspondence, which could be divided into the two categories: i) graph matching methods [5], which employ graph matching algorithms such as the Hungarian algorithm to build the cross-view correspondence; ii) instance identification methods [4], [18], which identify the cross-view within-category counterparts for the given samples and reestablish the correspondence between them. Although these methods achieve a promising performance, how to address both SP and VP in a unified framework is still less touched. Recently, SURE [4] formulates the solutions of SP and VP as a category-level identification task and proposes the first unified framework that simultaneously achieves robustness on both SP and VP.

The differences between the existing incomplete information restoration methods and our method are as follows. First, most existing methods focus on the sample-sample restoration paradigm, while CAMERA tackles both SP and VP through a sample-community framework, which could take advantage of community commonality and versatility. Second, to the best of our knowledge, this work could be the first attention-based framework in the field of incomplete information restoration. Such a framework successfully manipulates the attention between samples and communities that could be employed to impute missing samples and establish correspondences, showcasing its immense potential in data restoration. Thirdly, CAMERA could handle full sample-missing or view-unaligned problems without any paired samples, whereas most works rely on partially paired samples. More in-depth discussions with the sample-sample restoration paradigm are presented in Supplementary Material 2.

2.3 Learning with Noisy Correspondence

Noisy Correspondence (NC) refers to inherently irrelevant or relevant samples that are wrongly regarded as associated (*a.k.a*, false positive) or unassociated (*a.k.a*, false negative), which is first revealed and studied by [19], [20]. Learning with noisy correspondence aims to mitigate the negative impacts of false positive and false negative pairs, which has recently attracted increasing attention in visual instruction tuning [21], vision-language pre-training [22], [23], image-text matching [24], [25], graph matching [26], object re-identification [27], [28], and so on.

Both the aforementioned noisy correspondence problem and the view-unaligned problem studied in our work focus on the imperfect cross-view/modal correspondence issue. To be more specific, the former refers to incorrect correspondences, whereas the latter refers to missing correspondences. Beyond tackling the imperfect correspondence issue, our work further mitigates the negative impact of missing samples, leading to the so-called incomplete information problem.

3 Method

In this section, we introduce Community-Aware MultiviEw RepresentAtion learning with incomplete information (CAMERA), which improves incomplete information restoration performance by taking advantage of both community commonality and community versatility. In Section 3.1, we present a novel dual-stream network for deriving the sample- and community-level representation. In Section 3.2, we introduce the loss function that could embrace community versatility and further enhance community commonality. In Section 3.3, we offer the theoretical analysis of two community-level losses from the informative perspective. In Section 3.4, we elaborate on how CAMERA tackles both SP and VP under a unified MA-based framework.

3.1 Dual-stream network

In this section, we propose a dual-stream network, which endows view-specific sample-level representation with community commonality and learns community-level representation with aggregated sample information. For clarity, we denote the multi-view dataset as $\{X^v\}_{v=1}^V = \{x_1^v, x_2^v, \dots, x_N^v\}_{v=1}^V$, where $v \in [1, V]$, where V is the number of views and N is the number of samples.

As discussed in the Introduction, it is essential to guarantee community commonality by coupling the sample-level representation with the community-level representation. To this end, we propose modeling the relationship between samples and communities as the *mutual attention* A^v and utilizing it to integrate the sample- and community-level representations with each other. Mathematically,

$$A^v = \mathrm{Softmax}\left((W_S^v S^v)^T W_C^v C^v / \sqrt{d}\right), \tag{1}$$

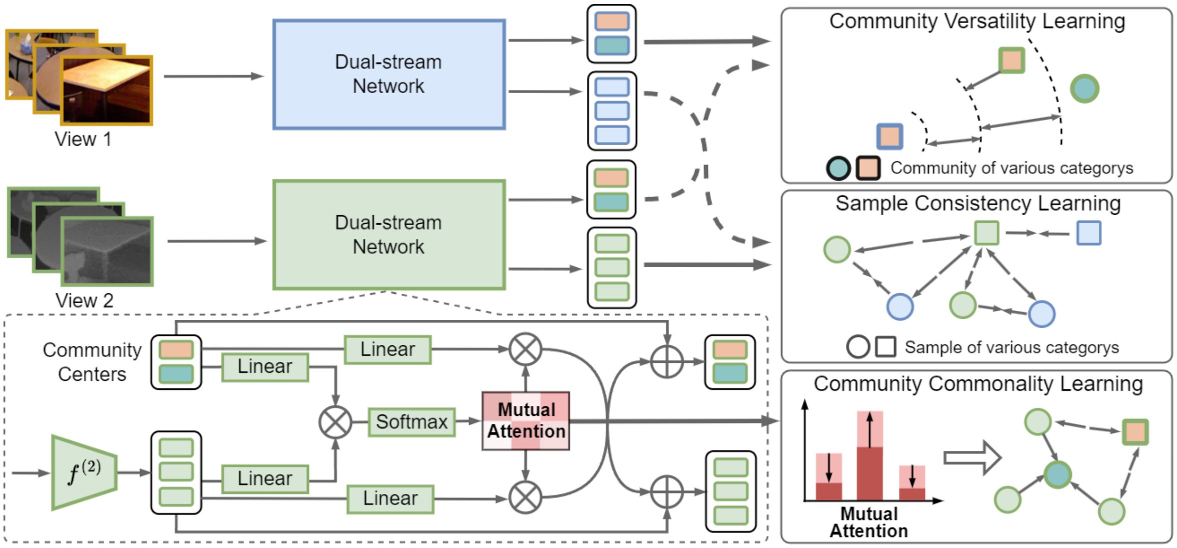

View 1
View 2
Dual-stream Network
Dual-stream Network
Community Versatility Learning
Community of various categorys
Sample Consistency Learning
Sample of various categorys
Community Commonality Learning
Community Centers
Linear
Linear
Softmax
Mutual Attention
Linear
Linear
$f^{(2)}$
Mutual Attention

Fig. 2. The framework of our method. Without loss of generality, we take the bi-view data as a showcase. We use the view-specific dual-stream encoder to extract representations for each view. Specifically, we explicitly learn a set of community centers and model the relationship between samples and community centers through a mutual attention mechanism. This mechanism is pivotal as it promotes the learning of both sample- and community-level representations, capturing the community commonality. To further enhance community commonality and preserve community versatility, we introduce a dual-level learning loss on the sample- and community-level representations with three joint objectives, *i.e.*, community versatility learning, sample consistency learning, and community commonality learning.

where $S^v \in \mathbb{R}^{d\times N}$ is the view-specific representation of X^v projected by the encoder $f^{(v)}$, $C^v \in \mathbb{R}^{d\times K} = \{c_1^v, c_2^v, \ldots, c_K^v\}$ is a set of learnable community centers with the random initialization, K is the category number, W_S^v, W_C^v are the projectors for S^v and C^v, respectively.

The mutual attention A^v plays a dual role, facilitating the learning process of sample- and community-level representation. At the sample level, the corresponding community center would be integrated into each sample by resorting to the sample-to-community attention, so that the sample-level representation Z^v could be induced. Formally,

$$Z^v = S^v + W'^v_C C^v \left(A^v\right)^T, \tag{2}$$

where W'^v_C is the projector for encoding C^v. Clearly, given a sample as the query and community centers as the gallery, the retrieved community center could be regarded as the corresponding community center to the sample. Accordingly, each sample would be pushed towards its corresponding community center, thereby endowing the within-community samples with commonality.

Similarly, at the community level, the within-community samples could be integrated into each community center through community-to-sample attention. As a result, the community-level representation U^v could be obtained. Formally,

$$U^v = C^v + W'^v_S S^v A^v, \tag{3}$$

where W'^v_S is the projector for encoding S^v. In other words, given a community center as the query and samples as the gallery, the retrieved samples could be regarded as the within-community samples. Therefore, such an operation inherently integrates within-community samples into the community center, which helps to formulate the community representation.

3.2 The Loss Function

Thanks to the dual-stream network, community commonality is endowed into the sample-level representation and the community-level representation is derived. To preserve the community versatility of the community-level representation and further enhance community commonality of the sample-level representation, we propose the following objective function,

$$\mathcal{L} = \mathcal{L}_{dl} + \lambda\mathcal{L}_{rec}, \tag{4}$$

where $\mathcal{L}_{dl}$ is the dual-level learning loss, $\mathcal{L}_{rec}$ is the reconstruction loss and λ is the trade-off parameter. In the following, we elaborate on each loss term one by one.

Following [29], the reconstruction loss $\mathcal{L}_{rec}$ is used to preserve sufficient information from the input data. To be specific, the loss is designed as follows,

$$\mathcal{L}_{rec} = \sum_{v=1}^{V}\sum_{i=1}^{N}\left\|x_i^v - g^{(v)}\left(z_i^v\right)\right\|_2^2, \tag{5}$$

where $g^{(v)}$ is the view-specific decoder of the v-th view.

The dual-level learning loss consists of the community- and sample-level losses. Formally,

$$\mathcal{L}_{dl} = \mathcal{L}_{ccl} + \mathcal{L}_{cvl} + \mathcal{L}_{scl}, \tag{6}$$

where the community-level commonality loss $\mathcal{L}_{ccl}$ imposes coherence to the view-specific within-community samples, the community-level versatility loss $\mathcal{L}_{cvl}$ is proposed to preserve the view diversity, and the sample-level consistency loss $\mathcal{L}_{scl}$ aims to help guarantee the consistency of sample-level representation across views. In the following, we will expound upon each of them individually.

3.2.1 *Community-level Commonality Learning*

With the help of the dual-stream network, we model the mutual attention A^v as an $N \times K$ matrix, where A^v_{ij} intrinsically corresponds to the probability of the i-th sample belonging to the j-th community. To further enhance community commonality of sample-level representation, we propose to learn more sharpened mutual attention. In other words, we expect each sample to be confidently assigned to a certain community. The aim could be achieved by the following community commonality loss,

$$\mathcal{L}_{ccl} = \sum_{v=1}^{V} \sum_{j=1}^{K} \left[A^v_{\cdot j} \log A^v_{\cdot j} - \frac{\alpha}{N} \sum_{i=1}^{N} A^v_{ij} \log A^v_{ij} \right], \quad (7)$$

where $A^v_{\cdot j} = \frac{1}{N} \sum_{i=1}^{N} A^v_{ij}$ and α is the balance weight which is fixed to 0.02 in all the experiments. The right part guarantees the sharpness of the mutual attention, while the left part prevents the situation where most samples are assigned to the same community (*i.e.*, trivial solution). Such a community commonality loss $\mathcal{L}_{ccl}$ encourages pushing samples to their corresponding community centers and being away from other centers, which would improve within-community compactness and between-community scatterness.

3.2.2 *Community-level Versatility Learning*

To embrace community versatility, we propose to learn the cross-view community-level representation in a non-monotonic manner, which avoids overemphasizing the cross-view community consistency. Formally, the proposed community versatility learning loss is defined as follows,

$$\mathcal{L}_{cvl} = \frac{1}{K} \sum_{v_1 \neq v_2}^{V} \sum_{i}^{K} \ell_i^{v_1,v_2},$$

$$\ell_i^{v_1,v_2} = \sum_{j \neq i}^{K} \left[\sigma(\beta) - \min\left(s\left(u_i^{v_1}, u_i^{v_2}\right), \beta\right) + s\left(u_i^{v_1}, u_j^{v_2}\right) \right]_+, \quad (8)$$

where $[\cdot]_+ = \max(\cdot, 0)$, $s(\cdot,\cdot)$ denotes the cosine similarity, β is the similarity bound for positives which is fixed as 0.7, $\sigma(\beta)$ is the margin, and $u_j^{v_2}$ is the negative community-level representation for a given $u_i^{v_1}$. Notably, the margin $\sigma(\beta) \propto \beta$ is designed to prevent the similarity of negatives from wrongly increasing due to the over-emphasized similarity bound of positives. For simplicity, we set $\sigma(\beta) = \beta^2$ in our implementation.

We design $\mathcal{L}_{cvl}$ for the following goal. Although the vanilla contrastive loss [30] could enforce the positive pairs to be similar, it would simply maximize the similarities of positive pairs and thus lose view-specific information. Clearly, such a solution would destroy the view complementary assumption [2], [31], thus being infeasible for MvRL. In contrast, $\mathcal{L}_{cvl}$ has an incentive to optimize the similarities of positive pairs to a bound β, which prevents the positive pairs from being identical. Besides, maintaining a margin $\sigma(\beta)$ guarantees the discrepancy between positives and negatives. Thanks to the community versatility learning loss, the community-level representation could embrace both similar and diverse patterns, in which the similar pattern helps to establish cross-view community correspondence, and the diverse pattern preserves the unique view-specific information for multi-view diversity.

3.2.3 *Sample-level Consistency Learning*

To facilitate the cross-view consistency of sample-level representation, we employ a sample-level contrastive loss to maximize the similarities between cross-view samples of the same instance, while simultaneously minimizing the similarities of samples from different instances. Formally,

$$\mathcal{L}_{scl} = \frac{1}{N} \sum_{v_1 \neq v_2}^{V} \sum_{i=1}^{N} \hat{\ell}_i^{v_1,v_2},$$

$$\hat{\ell}_i^{v_1,v_2} = -\log \frac{e^{s\left(z_i^{v_1}, z_i^{v_2}\right)/\tau_I}}{\sum_{j=1}^{N} \left[e^{s\left(z_i^{v_1}, z_j^{v_1}\right)/\tau_I} + e^{s\left(z_i^{v_1}, z_j^{v_2}\right)/\tau_I} \right]}, \quad (9)$$

where $\tau_I = 0.5$ is the temperature parameter.

3.3 Theoretical Analysis from the Informative Perspective

In this section, we conduct a theoretical analysis of the proposed community-level learning losses. In brief, we first define the community commonality and the community versatility from the informative perspective. Based on the definitions, we derive the lower bound of two community characteristics and propose a general objective function to embrace both community commonality and community versatility accordingly. The proposed community-level loss in Section. 3.2 is an effective implementation of the objective function.

In the information theory framework, $H(\cdot)$ denotes the entropy, $H(\cdot \mid \cdot)$ denotes the conditional entropy, $I(\cdot;\cdot)$ denotes the mutual information, and $I(\cdot;\cdot \mid \cdot)$ denotes conditional mutual information. It is worth noting that in real-world scenarios, the samples Z^v and community centers U^v are constructed from finite sets, with $|Z^v| = N$ and $|U^v| = K$. Accordingly, we treat Z^v and U^v as discrete random variables throughout the derivations, following the prior works [3], [32], [33]. In this case, the entropy $H(\cdot)$ and conditional entropy $H(\cdot \mid \cdot)$ are always non-negative. Besides, we assume that the sample-community relationship (*i.e.*, mutual attention A^v) is derived based on the community centers U^v, i.e., $H(A^v \mid U^v) = 0$. Based on the above notations and assumptions, we first give the definition of community commonality.

Definition 1 (Community Commonality)**.** *For the v-th view, community commonality is defined as the mutual information between within-view samples and the community centers, i.e., $I(Z^v; U^v)$.*

A large community commonality indicates that the samples are close to their corresponding community centers and are far from the other community centers.

Theorem 1. *The mutual information between mutual attention A^v and samples Z^v is the lower bound of community commonality $I(Z^v; U^v)$, i.e., $I(Z^v; U^v) \geq I(Z^v; A^v) = -H(A^v|Z^v) + H(A^v)$.*

The detailed proofs of Theorem 1 are presented in Supplementary Material 1.1. Theorem 1 indicates that the community commonality could be optimized by minimizing $H\left(A^v|Z^v\right)$ and maximizing $H\left(A^v\right)$, which could be implemented by the community commonality learning loss proposed in Eq. 7. Specifically, the first term $H\left(A^v|Z^v\right)$ could be rewritten as,

$$H\left(A^v|Z^v\right) = -\frac{1}{N}\sum_{i,j} A^v_{ij} \log A^v_{ij}. \tag{10}$$

As observed, the minimization of $H\left(A^v|Z^v\right)$ pushes each sample to its corresponding community center and away from the others, *i.e.*, CAMERA encourages improving within-community compactness and between-community scatterness.

The second term $H\left(A^v\right)$ is the entropy of mutual attention, namely,

$$H\left(A^v\right) = -\sum_{j} A^v_{\cdot j} \log A^v_{\cdot j}. \tag{11}$$

The maximization of $H\left(A^v\right)$ punishes too large or small communities in each view to prevent trivial solutions.

Definition 2 (Community Versatility)**.** *For the cross-view community centers U^{v_1} and U^{v_2}, the community versatility is defined as $H\left(U^{v_1}|U^{v_2}\right) + H\left(U^{v_2}|U^{v_1}\right)$.*

By the definition, we could preserve community versatility in the following way. For clarity, we elaborate on community versatility by taking bi-view data as a showcase.

Theorem 2. *The joint function of community commonality $I\left(Z^v;U^v\right)$ and mutual information between cross-view community centers $I\left(U^1;U^2\right)$ is the lower bound of community versatility, formally,*

$$\begin{aligned} &H\left(U^1|U^2\right) + H\left(U^2|U^1\right) \\ &\geq \sum_{v=1}^{2} I\left(Z^v;U^v\right) - 2I\left(U^1;U^2\right). \end{aligned} \tag{12}$$

The detailed proofs of Theorem 2 are presented in Supplementary Material 1.2. Drawing upon the theoretical insights, we could maximize the community commonality $I\left(Z^v;U^v\right)$ and minimize the mutual information $I\left(U^{v_1};U^{v_2}\right)$ of cross-view community representations to preserve the community versatility. However, motivated by the community concept in Fig. 1, the communities of the same category should be highly similar, which indicates that $I\left(U^{v_1};U^{v_2}\right)$ can not be simply minimized. To solve this problem, we propose to optimize $I\left(U^{v_1};U^{v_2}\right)$ to a margin, embracing both consistency and versatility. Accordingly, the joint objective function that helps to enhance community commonality and preserve community versatility is as follows,

$$\begin{aligned} &\max \sum_{v=1}^{V} \left(-H\left(A^v|Z^v\right) + H\left(A^v\right)\right) \\ &\text{s.t.} \min \sum_{v_1 \neq v_2}^{V} \left[M - I\left(U^{v_1};U^{v_2}\right)\right]_+ , \end{aligned} \tag{13}$$

where M is the margin. The optimization target in Eq. 13 could be achieved by community commonality learning loss based on Theorem 1.

Following, we try to optimize the constraint in Eq. 13. According to CPC [34] and DCP [3], the cross-view InfoNCE loss $\mathcal{L}_N$ could be regarded as the lower bound of cross-view mutual information, *i.e.*, $I\left(U^1,U^2\right) \geq \log N - \mathcal{L}_N$. However, InfoNCE loss only maximizes the mutual information while losing the view-specific information. To tackle this, we propose a novel triplet-form community-level loss $\mathcal{L}_{cvl}$. The community-level triplet loss could preserve the community versatility and proved to satisfy the constraint in Eq. 13 in Supplementary Material 1.3.

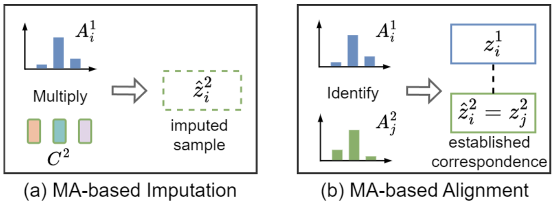


Fig. 3. MA-based Imputation and Alignment. (a) MA-based Imputation. The missing sample is imputed with the mutual attention inherited from the observed view and community centers in the missing view. (b) MA-based Alignment. The correspondence is established by confirming that the sample and its cross-view counterpart embrace consistent mutual attention.

3.4 MA-based Incomplete Information Restoration

In this section, we formally define the incomplete information and present the details of our MA-based framework for incomplete information restoration

Definition 3 (Incomplete Information)**.** *The dataset $\left\{X^v\right\}_{v=1}^{V}$ is with incomplete information when a portion of the dataset, i.e., $\left\{Q^v\right\}_{v=1}^{V} = \left\{q_1^v, q_2^v, \ldots, q_{N_q}^v\right\}_{v=1}^{V}$ is contaminated with SP, VP, or both of SP and VP, where N_q is the number of instances with incomplete information. Specifically, $\left\{Q^v\right\}_{v=1}^{V}$ is sample-missing when*

$$1 \leq \sum_{v=1}^{V} Obs\left(q_i^v\right) < V,\ \forall i \in [1, N_q] \tag{14}$$

where $Obs(\cdot)$ is an indicator function evaluating to 1 for the observed samples. While $\left\{Q^v\right\}_{v=1}^{V}$ is view-unaligned when

$$\sum_{v_1}^{V} \sum_{v_2 \neq v_1}^{V} Cor\left(q_i^{v_1}, q_i^{v_2}\right) < V\left(V-1\right),\ \forall i \in [1, N_q] \tag{15}$$

where $Cor(\cdot,\cdot)$ is an indicator function evaluating to 1, i.f.f., samples belong to the same instance.

To implement the incomplete information restoration, we design two solutions toward SP and VP based on the mutual attention mechanism in the inference stage, *i.e.*, MA-based Imputation (MAI) and MA-based Alignment (MAA), respectively. For clarity, in the following, we elaborate on the MAI and MAA operations by taking bi-view data as a showcase.

MA-based Imputation (MAI). As shown in Fig. 3(a), with the observed sample x_i^1 in view 1, the missing sample in view 2 could be imputed with the help of community centers C^2 from the missing view and the mutual attention of x_i^1. Formally,

$$\hat{z}_i^2 = s_i^1 + W'^2_C C^2 \left(A_i^1\right)^T, \tag{16}$$

where A_i^1 is the mutual attention of x_i^1 and $\hat{z}_i^2$ is the imputed sample representation in view 2. Notably, the representation of x_i^1 is also utilized in the imputation for maintaining the cross-view consistency.

MA-based Alignment (MAA). As shown in Fig. 3(b), given x_i^1 in view 1 and its counterpart sample in view 2, their correspondence is confirmed by resorting to mutual attention. Specifically, the two samples would be regarded as paired, $i.f.f.$,

$$\arg\max A_i^1 = \arg\max A_i^2. \tag{17}$$

Otherwise, the two samples are unaligned, and a new corresponding sample x_j^2 from view 2 would be derived for x_i^1. The representation of the newly-corresponded sample x_j^2 could be defined as

$$\begin{gathered} \hat{z}_i^2 = s_j^2 + {W'}_C^2 C^2 \left(A_j^2\right)^T \\ \text{s.t. } s_j^2 \in \mathcal{N}^2\left(s_i^1\right),\ \arg\max A_i^1 = \arg\max A_j^2, \end{gathered} \tag{18}$$

where $\mathcal{N}^2\left(s_i^1\right)$ indicates the nearest cross-view neighbors of the sample representation s_i^1. In our implementation, we seek the nearest neighbor in $\mathcal{N}^2\left(s_i^1\right)$ as x_j^2 which satisfies the constraints in Eq. 18.

For cross-view samples of the same instance, the idea behind the MA-based framework is that their mutual attentions are expected to be consistent across different views. Thanks to the explicit cross-view community correspondence established by community versatility, mutual attention (sample-community relationship) could be employed to implement the MA-based framework. Such a framework would take advantage of community commonality and versatility, which benefits incomplete information restoration. Specifically, our framework improves the compactness of view-specific clusters and preserves multi-view diversity by incorporating the community center with community- and view-specific information into the sample, respectively. Clearly, both advantages are indispensable in MvRL, which indicates that the MA-based framework helps to improve the representation after restoration.

4 Experiments

In this section, we evaluate the proposed CAMERA on three different multi-view learning tasks, including clustering, classification, and human action recognition. The section is structured as follows. In Section 4.1, we introduce the settings of the experiments. In Section 4.2, we elaborate on the network architectures and the implementation details of CAMERA. In Sections 4.3-4.5, we conduct experiments on various multi-view tasks to verify the effectiveness of the proposed CAMERA. In Section 4.6, we investigate the robustness of CAMERA through a series of parameter analyses, quantitative analyses, ablation studies, and visualization analyses.

4.1 Experimental Settings

We conduct experiments on the seven widely-used datasets, including Scene-15 [36], Reuters [37], NoisyMNIST [11], CUB [38], LandUse-21 [39], UWA [40] and DHA [41]. The details of these datasets are presented in Supplementary Material 3.

Following [4], [18], for the multi-view clustering and classification tasks, we adopt the below setting. For the sample-missing setting, we randomly remove one view of N_q instances to generate the data with missing samples, and the missing rate is defined as $\eta_{SP} = N_q/N$. For the view-unaligned setting, we randomly shuffle N_q samples of the unaligned view to remove the correspondences, and the unaligned rate is defined as $\eta_{VP} = N_q/N$.

4.2 Network Architectures and Implementation Details

Following [3], [18], CAMERA employs the architectures of convolutional auto-encoder for multi-view image datasets (*i.e.*, NoisyMNIST) and fully-connected auto-encoder for other datasets. For evaluation, all view-specific representations Z^v are concatenated as the fusion representation following most existing multi-view learning methods [5], [19]. Then, we perform k-means [9] on the fusion representation to obtain the clustering results, while SVM [42] is employed to obtain the classification results. As for the human action recognition task, we employ the classifier upon the fusion representation and view-specific representation to obtain the results. For all experiments, we repeat each method with five different random initializations and report the mean results for fair comparisons.

The proposed CAMERA is implemented in PyTorch 1.11.0 and the experiments are carried out on an NVIDIA 3090 GPU. We utilize Adam optimizer for all the datasets and the batch size is set to 256. For the clustering and classification tasks, the model is trained for 150 epochs with an initial learning rate of 0.001. In contrast, the model is trained within 300 epochs with an initial learning rate of 0.0003 for the human action recognition task. The balance weight α in Eq. 7 and the bound β in Eq. 8 are fixed to 0.02 and 0.7, respectively. The balance weight λ is fixed to 10 in all the datasets except for the multi-view image dataset (NoisyMNIST) and multi-view RGBD dataset (UWA, DHA), whose λ is fixed to 1. In practice, in the first 50 epochs, the model is warmed up with losses except for the community versatility learning loss since it would falsely pull between-community samples at the early stage. To establish the correspondence between cross-view communities, CAMERA first utilizes mutual attention to obtain the community assignment of the samples and then performs the Hungarian algorithm on the paired samples to establish community correspondence across views. Finally, the model is trained with the overall loss (Eq. 4) during the rest training time.

4.3 Comparisons on Multi-view Clustering

In this section, we carry out experiments on the multi-view clustering (MvC) task and compare our CAMERA with 16 state-of-the-art multi-view clustering baselines. The baseline methods could be divided into six kinds: i) the vanilla MvC methods including DCCA [10], DCCAE [11], BMVC [16] and AE2-Nets [12]; ii) the MvC methods against partial SP (Eq. 14, $N_q < N$) including PMVC [14], EERIMVC [15], DCP [3], DSIMVC [1], and ProImp [35]; iii) the MvC methods against partial VP (Eq. 15, $N_q < N$) including PVC [5] and MVCLN [19]; iv) the generalized MvC methods against both partial SP and partial VP including SURE [4] and SMILE [18]; v) the MvC methods against full SP (Eq. 14,

TABLE 1
The multi-view clustering performance comparisons on five widely-used benchmarks. The best and second best results are denoted in **bold** and underline, respectively.

Setting	Method	Scene-15			Reteurs			NoisyMNIST			LandUse-21			CUB		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Missing	DCCA [10]	28.8	28.4	13.2	45.8	26.1	18.0	61.8	60.6	37.7	14.1	20.0	3.4	44.2	43.3	26.7
	DCCAE [11]	29.0	29.1	12.9	47.0	28.0	14.5	65.4	62.9	38.3	14.9	20.9	3.7	42.3	40.9	25.5
	BMVC [16]	32.5	30.9	11.6	32.1	7.0	2.9	30.7	19.2	10.6	18.8	18.7	3.7	29.8	20.3	6.4
	AE-Nets [12]	22.4	23.4	9.6	29.1	7.6	4.8	29.9	23.8	11.8	19.2	23.0	5.8	35.9	32.0	15.9
	PMVC [14]	25.5	25.4	11.3	29.3	7.4	4.4	33.1	25.5	14.6	20.0	23.6	8.0	57.7	54.4	38.3
	PVC [5]	27.0	23.5	10.6	20.7	5.3	3.8	16.4	6.7	2.3	21.3	23.1	8.1	39.0	40.5	20.9
	EERIMVC [15]	31.5	31.1	14.8	29.8	12.0	4.2	55.6	45.9	36.8	22.1	25.2	9.1	68.7	63.9	53.8
	MVCLN [19]	31.4	29.5	13.9	39.3	18.4	14.3	53.8	50.6	28.5	22.1	25.2	9.1	45.2	40.8	21.9
	SURE [4]	39.6	41.6	23.5	47.2	30.9	23.3	93.0	85.4	85.9	23.1	28.6	10.6	58.3	50.4	37.4
	DCP [3]	39.5	42.4	23.5	34.6	17.5	2.9	80.0	75.2	70.7	22.2	27.0	10.4	53.7	65.5	47.3
	DSIMVC [1]	30.6	35.5	17.2	39.9	19.6	17.1	55.8	55.1	43.0	18.6	18.0	5.7	54.4	52.4	35.2
	ProImp [35]	41.6	42.9	25.3	51.9	35.5	28.5	94.9	87.4	89.1	22.4	26.6	9.9	73.3	66.4	54.8
	SMILE [18]	41.5	41.3	25.3	39.4	30.0	23.5	96.8	91.7	93.0	24.5	28.3	11.4	69.5	66.7	54.9
	Ours	**44.9**	**44.4**	**26.9**	**54.4**	**35.9**	**30.3**	**98.3**	**95.1**	**96.3**	**27.2**	**31.9**	**13.4**	**74.1**	**68.1**	**56.3**
Unaligned	DCCA [10]	34.3	36.6	18.8	39.7	13.8	14.4	34.5	29.8	17.9	20.5	22.5	7.5	15.9	3.3	0.1
	DCCAE [11]	33.6	36.6	18.5	41.4	12.8	14.4	27.6	19.5	10.0	18.2	18.9	5.6	15.8	2.8	0.2
	BMVC [16]	29.5	29.9	14.8	38.2	11.6	12.1	28.5	24.7	14.2	13.8	11.8	2.9	16.0	3.4	0.2
	AE-Nets [12]	36.8	36.6	20.2	35.5	10.6	8.1	38.3	34.3	22.0	12.0	8.7	1.5	14.5	2.6	0.3
	PMVC [14]	30.1	27.8	14.4	24.6	3.6	1.9	31.9	21.4	13.0	22.2	25.2	9.4	15.8	3.0	0.0
	PVC [5]	37.9	39.1	20.6	42.1	20.4	17.0	81.8	82.3	82.0	23.6	30.0	9.9	50.2	56.3	38.6
	EERIMVC [15]	25.0	21.3	10.3	39.9	14.9	14.0	46.8	29.6	23.9	22.8	22.3	9.7	15.8	2.9	0.0
	MVCLN [19]	38.5	39.9	24.3	50.2	30.7	24.9	91.1	84.2	83.6	25.0	27.9	11.6	58.2	55.2	40.8
	SURE [4]	40.3	40.3	23.1	50.0	29.5	24.6	95.2	88.2	89.7	24.9	28.6	11.8	64.5	62.0	47.9
	DCP [3]	28.1	29.4	12.5	36.2	9.9	7.0	32.3	28.0	9.4	21.2	23.2	8.3	35.4	30.7	8.1
	DSIMVC [1]	24.4	26.1	11.0	41.5	21.4	18.7	34.6	24.0	16.8	17.3	17.3	4.8	30.4	25.4	11.8
	SMILE [18]	41.3	41.1	24.7	40.9	30.4	24.5	97.9	94.2	95.4	26.6	28.8	12.8	71.1	70.4	58.2
	Ours	**44.8**	**44.1**	**27.1**	**54.6**	**32.5**	**27.9**	**98.2**	**95.0**	**96.0**	**27.9**	**33.9**	**14.2**	**75.7**	**71.1**	**60.5**
Complete	DCCA [10]	36.6	39.2	21.0	48.0	26.6	12.7	89.6	88.3	84.0	15.5	23.2	4.4	55.6	56.1	43.2
	DCCAE [11]	34.6	39.0	19.7	42.0	20.3	8.5	78.0	81.2	68.2	15.6	24.4	4.4	55.3	58.7	45.1
	BMVC [16]	40.5	41.2	24.1	42.4	21.9	15.1	88.3	77.0	76.6	25.3	38.6	11.4	66.2	61.7	48.7
	AE-Nets [12]	37.2	40.5	22.2	42.4	19.8	14.9	42.1	43.4	30.4	24.8	30.4	10.4	48.8	46.7	30.5
	PMVC [14]	30.8	31.1	15.0	32.5	11.1	7.5	41.1	36.4	24.5	25.0	31.1	12.2	64.5	70.3	53.1
	PVC [5]	38.0	39.8	21.1	47.7	24.4	17.7	87.1	92.8	93.1	25.2	30.5	11.7	59.7	65.3	51.6
	EERIMVC [15]	39.6	39.0	22.1	33.2	14.3	3.9	65.7	57.6	51.3	24.9	29.6	12.2	74.0	73.1	62.4
	MVCLN [19]	40.5	41.8	24.8	49.1	30.7	26.4	97.3	94.2	95.3	25.2	28.2	11.8	59.7	56.5	42.5
	SURE [4]	41.0	43.2	25.0	49.1	29.9	23.6	98.4	95.4	96.5	25.1	28.3	10.9	58.0	59.3	45.2
	DCP [3]	41.1	44.7	24.8	36.2	18.9	4.8	89.1	88.9	85.5	25.6	31.7	13.1	63.6	70.2	53.9
	DSIMVC [1]	31.7	35.6	17.2	43.2	23.3	19.0	61.0	58.1	46.7	18.1	18.6	5.6	58.5	56.3	39.9
	ProImp [35]	43.6	45.0	26.8	**56.5**	**39.4**	**32.8**	99.2	97.5	98.2	23.7	27.9	10.8	80.7	75.1	65.4
	SMILE [18]	44.4	44.6	27.4	42.5	32.9	26.2	99.3	97.8	98.4	26.7	29.1	13.1	74.7	75.5	64.5
	Ours	**45.5**	**46.3**	**28.2**	55.0	34.8	29.5	**99.9**	**99.6**	**99.7**	**28.6**	**35.4**	**15.1**	**81.4**	**75.6**	**66.6**

$N_q = N$) including DM2C [43]; vi) the MvC methods against full VP including GWMAC [44] and MVC-UM [45]. In the experiments, we train all the baseline methods with their suggested parameters. Moreover, for the vanilla MvC methods that cannot handle SP or VP, we employ the following two prepossessing steps for fair comparisons:

- For baselines that can't handle SP, we impute the missing samples by using the mean of the observed samples in the missing view. After that, the baselines are carried out on the imputed data.
- For baselines that can't handle VP, we first employ PCA [46] to project the data into the latent space, then perform the Hungarian algorithm to establish the correspondence. After that, the baselines are carried out on the re-aligned data.

Three widely-used metrics are used to evaluate the performance of the clustering task, namely, Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Higher values of these metrics signify superior clustering performance.

In the settings of partial SP (missing rate $\eta_{SP} = 50\%$, denoted as "Missing"), partial VP (unaligned rate $\eta_{VP} = 50\%$, denoted as "Unaligned"), and complete information (both $\eta_{SP} = 0\%$ and $\eta_{VP} = 0\%$, denoted as "Complete"), we compare CAMERA with 13 multi-view clustering baselines in Tab. 1, where one could see that: i) CAMERA exhibits a remarkable performance superiority over state-of-the-art methods across most datasets. For example, on the Scene-15 dataset, CAMERA outperforms the best baseline with a relative performance improvement of 7.9% (44.9% v.s. 41.6%) and 8.5% (44.8% v.s. 41.3%) in terms of ACC under "Missing" and "Unaligned" settings, respectively; ii) CAMERA achieves advanced performance in the "Complete" setting, showing the benefits of embracing community commonality and versatility.

Besides, to verify the robustness of our CAMERA, we evaluate it in a more challenging setting, namely, "Unpaired" refers to the data simultaneously contaminated with both SP and VP. To be more specific, in the setting, we randomly select N_q instances, and half of these instances are specified as the "Missing" setting while the others are specified as the "Unaligned" setting. For extensive evaluations, we vary the rates of Missing/Unaligned/Unpaired from 0% to 100% with an interval of 10%. For comparisons, we choose the three MvC methods (GWMAC, MVC-UM, and DM2C) against either full SP or full VP, along with the most competitive methods (PVC, MVCLN, DCP, SURE, ProIMP, SMILE) in Tab. 1 as the baselines. Notably, in the setting of full SP or full VP, CAMERA regards samples and their cross-view nearest neighbors as the pairwise samples and constructs the cross-view community correspondence as there are no prior paired samples. From

TABLE 2
The multi-view classification performance comparisons on five benchmarks. The best and second best results are denoted in **bold** and underline, respectively.

Setting	Method	Scene-15		Reteurs		NoisyMNIST		LandUse-21		CUB	
		ACC	Pre	ACC	Pre	ACC	Pre	ACC	Pre	ACC	Pre
Missing	DCP [3]	44.9	42.2	70.7	70.8	87.8	87.6	25.3	27.0	57.0	69.4
	SURE [4]	51.4	49.2	80.0	77.2	94.4	94.4	29.8	31.6	54.2	52.5
	ProImp [35]	54.1	51.0	82.2	81.6	94.9	94.8	31.2	30.8	77.5	78.2
	SMILE [18]	52.4	50.1	61.3	58.3	96.6	96.6	45.9	44.7	77.5	78.0
	Ours	**66.7**	**65.6**	**84.7**	**84.6**	**98.4**	**98.4**	**51.9**	**52.8**	**84.2**	**84.2**
Unaligned	PVC [5]	50.3	49.0	79.5	76.0	80.4	80.3	41.1	42.0	75.5	75.0
	MVCLN [19]	48.2	46.0	78.6	76.7	95.7	95.7	35.4	36.7	69.4	69.0
	SURE [4]	51.1	49.9	79.7	76.5	96.1	96.1	34.0	35.3	54.6	50.2
	SMILE [18]	50.9	49.9	65.3	62.7	97.2	97.2	51.5	52.3	83.1	82.7
	Ours	**69.1**	**68.0**	**84.0**	**83.8**	**97.4**	**97.4**	**56.1**	**56.8**	**83.7**	**84.6**
Complete	PVC [5]	50.7	49.0	81.9	82.0	88.9	88.7	41.5	42.7	81.6	82.7
	MVCLN [19]	51.1	51.7	82.1	81.7	98.6	98.7	38.9	41.7	63.9	65.7
	DCP [3]	47.4	46.6	73.3	71.2	93.3	93.4	28.4	28.4	81.2	82.8
	SURE [4]	50.7	51.3	82.1	82.3	98.6	98.6	38.0	40.8	68.3	61.9
	ProImp [35]	56.6	54.6	83.4	83.2	99.2	99.0	38.6	38.8	82.5	83.0
	SMILE [18]	52.4	54.1	72.0	69.7	99.3	99.2	57.8	59.0	86.1	86.8
	Ours	**73.7**	**73.0**	**86.2**	**85.8**	**99.8**	**100.0**	**63.0**	**63.6**	**88.8**	**89.4**

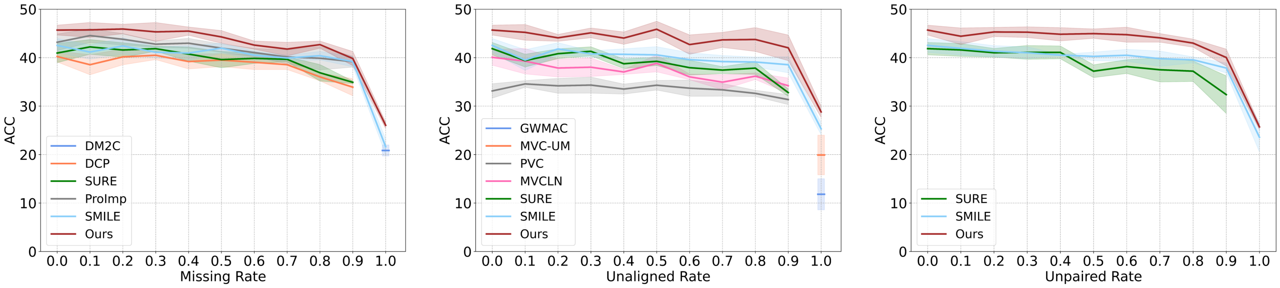


Fig. 4. Multi-view clustering performance analysis on the Scene-15 dataset with different Missing/Unaligned/Unpaired rates.

the results in Fig. 4, one could have the following observations: i) our CAMERA significantly outperforms baselines under all Missing/Unaligned/Unpaired rates, which demonstrates the robustness of our MA-based framework against all incomplete information cases; ii) when the Missing/Unaligned/Unpaired rate reaches 100%, most baselines cannot handle this situation. In contrast, the proposed CAMERA still achieves a great performance, which proves the effectiveness of CAMERA against multi-view data with full incomplete information.

4.4 Comparisons on Multi-view Classification

In this section, we carry out experiments on the multi-view classification task, comparing CAMERA with the 6 most competitive baselines in Tab. 1. For a comprehensive evaluation, we employ two widely-used classification metrics: Accuracy and Precision. Higher values of these metrics signify superior classification performance. Following [3], [19], the dataset is separated into train and test sets with a ratio of $8:2$.

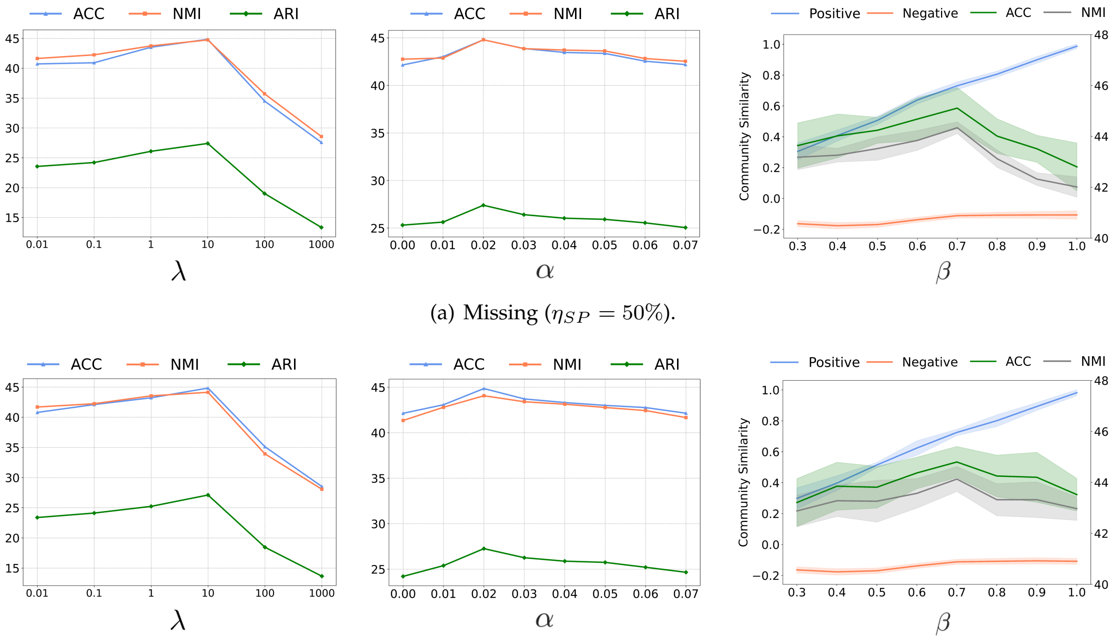
As shown in Tab. 2, CAMERA outperforms all the baselines in the Missing/Unaligned/Complete setting, which demonstrates the effectiveness of CAMERA on the classification task.

TABLE 3
The performance comparisons on the multi-view human action recognition task under the UWA dataset. In the table, RGB (R), Depth (D) and $\mathrm{R}+\mathrm{D}$ denote the RGB view, the depth view and the fusion of them, respectively. $\mathrm{R}\rightarrow\mathrm{D}$ indicates that the view D is generated by the view R, $\mathrm{D}\rightarrow\mathrm{R}$ is defined similarly as $\mathrm{R}\rightarrow\mathrm{D}$. "-" indicates that the baselines cannot handle the setting.

Method	RGB	$\mathrm{R}\rightarrow\mathrm{D}$	Depth	$\mathrm{D}\rightarrow\mathrm{R}$	$\mathrm{R}+\mathrm{D}$
LIBSVM [47]	69.4	68.5	34.9	34.3	72.7
VLAD [48]	71.5	-	-	-	-
TSN [49]	71.0	-	-	-	-
WDMM [50]	-	-	46.6	-	-
AMGL [51]	69.2	71.5	39.9	36.0	68.5
MLAN [52]	67.2	67.2	33.3	33.6	66.6
GMVAR [53]	-	73.5	-	50.4	76.3
GVCA [54]	-	-	-	-	77.1
DCP [3]	79.9	79.7	50.4	50.2	79.0
Ours	**80.8**	**80.3**	**51.8**	**50.9**	**83.4**

4.5 Comparisons on Multi-view Human Action Recognition

In this section, we carry out experiments on the multi-view human action recognition task and compare CAMERA with 9 baselines, including LIBSVM [47], VLAD [48], TSN [49], WDMM [50], AMGL [51], MLAN [52], GMVAR [53], GVCA [54] and DCP [3]. Following [3], we extract the RGB features and depth features by TSN and WDMM, respectively. In brief, the RGB view is selected from three snippets from each video in the DHA and UWA datasets and extracted by ResNet-101, and the depth view is extracted by the same scheme following [53], [54]. Following the widely-used protocol of the human action recognition tasks [53], [54], we additionally add view-specific, view-integration classifiers upon our CAMERA model and adopt the cross-entropy loss for training. In the experiment, we use 50% of the instances as the train set and the rest instances as the test set. All methods including our CAMERA are trained



(a) Missing ($\eta_{SP} = 50\%$).

(b) Unaligned ($\eta_{VP} = 50\%$).

Fig. 5. Parameter analysis on the Scene-15 dataset under the multi-view clustering task, *w.r.t.*, the trade-off parameter λ, the balance weight α in community commonality learning loss, the bound β in community versatility learning loss.

using both the RGB and Depth views while being evaluated under the following settings. "RGB", "Depth", and "R+D" means that the methods are evaluated by only using the off-the-shelf RGB view, depth view, and a combination of them, respectively. In contrast, under the setting of "$\mathrm{R} \rightarrow \mathrm{D}$", the methods first recover the depth view and then are evaluated based on the off-the-shelf RGB view and the generated view. Similarly, the methods under the "$\mathrm{D} \rightarrow \mathrm{R}$" setting first generate the RGB view and then are evaluated by the RGB and depth views accordingly.

From the results in Tab. 3, one could observe that CAMERA significantly outperforms all baselines, which illustrates the effectiveness of CAMERA. The performance superiority on the setting of "RGB" and "Depth", "$\mathrm{R} \rightarrow \mathrm{D}$" and "$\mathrm{D} \rightarrow \mathrm{R}$", "$\mathrm{R} + \mathrm{D}$" could be attributed to the effects of our commonality learning module, MA-based imputation module, versatile learning module, respectively.

TABLE 4
Multi-view clustering performance comparisons on community commonality with different incomplete information restoration paradigms on the Scene-15 dataset. * indicates without the aggregation operation (Eq. 2).

Setting	Method	Silhouette Score	ACC
Missing	DCP	0.56	39.5
	SURE	1.48	39.6
	SMILE	1.81	41.5
	CAMERA*	1.34	43.0
	CAMERA	**2.00**	**44.9**
Unaligned	PVC	0.69	37.9
	SURE	1.37	40.3
	SMILE	1.67	41.3
	CAMERA*	1.58	43.1
	CAMERA	**2.01**	**44.8**
Complete	PVC	1.28	38.0
	DCP	1.98	41.1
	SURE	2.56	41.0
	SMILE	2.45	44.4
	CAMERA*	2.02	44.7
	CAMERA	**2.86**	**45.5**

4.6 Parameter Analysis and Ablation Studies

In this section, we carry out a series of parameter analysis and ablation studies to investigate the robustness of CAMERA and explore the effect of community versatility and community commonality. Unless otherwise stated, all the experiments are conducted on the Scene-15 dataset under the multi-view clustering task.

4.6.1 Influence of Hyper-parameters

We investigate the influence of the hyper-parameters including the trade-off parameter λ (Eq. 4), the balance weight α (Eq. 7) in community commonality learning loss, and the bound β (Eq. 8) in the community versatility learning loss under the "Missing" and "Unaligned" setting. From the results in the left panel of Fig. 5, one could observe that a good choice of λ would improve the performance, and the best performance is achieved when $\lambda = 10$. To investigate the influence of α, we vary α within the range of 0 to 0.07 with an interval of 0.01. As shown in the middle panel of Fig. 5, one could observe that CAMERA is not sensitive to the choice of α unless $\alpha \neq 0$ where the attention has a tendency to be uniform thus losing the community commonality. To explore the community versatility, we vary β within the range of 0.3 to 1.0 with an interval of 0.1. As depicted in the right panel of Fig. 5, CAMERA performs stably within the range of $[0.5, 0.8]$ and achieves the best performance when $\beta = 0.7$. In other words, the over-small similarity of positives would degrade the cross-view consistency, while the over-large similarity of positives would lose the view complementary, and thus either of them would lead to a

TABLE 5
Ablation study of four loss terms on the Scene-15 dataset, where " ✓ " denotes the loss is adopted.

Setting	$\mathcal{L}_{ccl}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{scl}$	$\mathcal{L}_{cvl}$	ACC	NMI	ARI
Missing	✓				29.0	29.5	13.9
	✓	✓			30.0	30.1	14.6
	✓		✓		40.1	41.6	23.2
	✓			✓	20.7	18.7	7.4
	✓	✓	✓		42.6	42.5	25.4
	✓	✓		✓	28.6	29.6	14.2
	✓		✓	✓	42.4	42.0	24.7
	✓	✓	✓	✓	**44.9**	**44.4**	**26.9**
Unaligned	✓				22.3	19.7	8.6
	✓	✓			26.9	23.6	11.3
	✓		✓		40.0	41.5	23.1
	✓			✓	18.6	15.4	6.2
	✓	✓	✓		42.9	42.7	25.8
	✓	✓		✓	26.7	23.1	10.9
	✓		✓	✓	40.9	41.7	23.7
	✓	✓	✓	✓	**44.8**	**44.1**	**27.1**

sub-optimal performance.

4.6.2 Importance of the Community Commonality

To demonstrate the significance of embracing the community commonality in multi-view learning, we quantify the community commonality with the help of the silhouette score. In brief, the silhouette score could indicate within-cluster compactness and between-cluster scatterness. Specifically, we compute the silhouette score on the fusion representation after the restoration, and a larger value of the silhouette score indicates better community commonality. The results in Tab. 4 illustrate that our CAMERA could enhance the community commonality and thus improve the performance.

4.6.3 Ablation Studies

To explore the effectiveness of each loss term in CAMERA, we carry out ablation experiments on the four loss terms. According to the results shown in Tab. 5, one could observe that each loss term plays an inseparable role in the optimization. It is worth noting that the community versatility learning loss $\mathcal{L}_{cvl}$ relies on the well-established representation and attention brought by the sample consistency learning loss $\mathcal{L}_{scl}$. In other words, simply employing $\mathcal{L}_{cvl}$ on the exhausted representation and attention would falsely pull between-community samples, thus degrading the performance.

4.6.4 Effect of the Community Formulation

As discussed in the Introduction, our dual-stream network would derive a favorable formulation of community, thus benefiting the community characteristics. To prove the superiority of the dual-stream network, we further explore the following variants of community formulations:

- Learnable: build a set of learnable community centers and treat them as communities.
- k-means: employ k-means at each epoch to derive centroids and treat them as the proxy of the communities.
- k-means†: employ k-means at each epoch to derive the community assignment of samples and then aggregate the within-community samples to formulate the community.
- CAMERA: build a set of learnable community centers and then utilize the mutual attention to aggregate the within-community samples as the community.

TABLE 6
Different variants to formulate the community on the Scene-15 dataset. † indicates the operation that aggregates samples to formulate the community.

Setting	Strategy	ACC	NMI	ARI
Missing	k-means	42.8	43.0	25.6
	k-means†	43.0	43.9	25.8
	Learnable	42.9	42.5	25.9
	CAMERA	**44.9**	**44.4**	**26.9**
Unaligned	k-means	42.9	43.2	25.7
	k-means†	43.2	43.7	26.0
	Learnable	43.0	43.0	25.4
	CAMERA	**44.8**	**44.1**	**27.1**

TABLE 7
Ablation study of different data imputation and alignment strategies on the Scene-15 dataset.

Strategy	Imputation			Alignment		
	ACC	NMI	ARI	ACC	NMI	ARI
Sample Only	32.1	36.4	18.8	38.6	41.5	23.0
Community Only	32.6	35.0	18.0	33.8	39.5	20.2
CAMERA	**44.9**	**44.4**	**26.9**	**44.8**	**44.1**	**27.1**

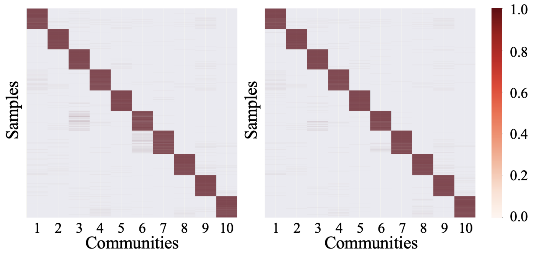
As shown in Tab. 6, one could have the following conclusions: i) treating the learnable community centers as communities would degrade the performance because such community centers don't explicitly aggregate the sample representation and thus might ineffectively represent the community; ii) although similar aggregation effect to the dual-stream network would be achieved, employing k-means to formulate the community would suffer from the outdated problem and the loss of community evolution information, which would degrade the performance. In contrast, the proposed dual-stream network builds a set of learnable community centers to preserve historical information and performs soft aggregation at each iteration for updating the community. Thanks to the dual-stream network, our strategy tackles these problems and thus achieves better performance.

4.6.5 Variants of Restoration Strategy

We conduct analytic experiments to investigate the various imputation and alignment strategies. Specifically,

- Imputation Strategies: "Sample Only" indicates recovering the missing sample using the observed sample, *i.e.*, $\hat{z}_i^2 = 2s_i^1$. "Community Only" indicates that recovering the missing sample using the community centers of the missing view, *i.e.*, $\hat{z}_i^2 = 2{W'}_C^2 C^2 \left(A_i^1\right)^T$.
- Alignment Strategies: given x_i^1 as the anchor sample, the new corresponding sample x_j^2 could be derived by nearest neighbor search based on the sample or community representations. Specifically, in the "Sample Only" setting, $j = \arg\min_j \|s_i^1 - s_j^2\|$, while in the "Community Only" setting, $j = \arg\min_j \|{W'}_C^1 C^1 (A_i^1)^T - {W'}_C^2 C^2 (A_j^2)^T\|$.

From the results in Table 7, one could observe that employing either the sample representation or the community representation would lose community versatility or community commonality, thus remarkably degrading the performance.



(a) View 1 (ACC = 95.8)
(b) View 2 (ACC = 96.2)

Fig. 6. Mutual Attention visualization on the NoisyMNIST dataset under Unpaired setting.

In contrast, the proposed CAMERA would take advantage of both community commonality and community versatility, thus boosting the performance.

4.6.6 Visualization on Mutual Attention

As discussed in Section 3.4, we employ mutual attention for data imputation and alignment. To verify the effectiveness of the attention-based incomplete information restoration strategy, we conduct the analytic experiments by visualizing the mutual attention. From the visualization results in Fig. 6, one could have the following conclusions. On the one hand, the mutual attention could distinguish samples from different categories, demonstrating that it indeed captures the intrinsic semantics of the samples. On the other hand, the mutual attention is relatively consistent across views, which supports the proposed attention-based imputation and alignment paradigm.

5 Conclusion

In this paper, we propose a robust MvRL method from a novel community perspective. Motivated by two concepts in sociology, we reveal that community commonality and versatility would benefit incomplete information restoration. To implement the two community characteristics, the proposed CAMERA employs a dual-stream network and a novel objective function. Moreover, we propose a novel MA-based framework to restore incomplete information by taking advantage of community characteristics. Experiments and theoretical analysis demonstrate that CAMERA captures the community commonality and versatility and thus boosts the performance of MvRL with incomplete information.

References

[1] H. Tang and Y. Liu, "Deep safe incomplete multi-view clustering: Theorem and algorithm," in *ICML*, 2022.

[2] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[3] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, "Dual contrastive prediction for incomplete multi-view representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] M. Yang, Y. Li, P. Hu, J. Bai, J. C. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[5] Z. Huang, P. Hu, J. T. Zhou, J. Lv, and X. Peng, "Partially view-aligned clustering," *NeurIPS*, 2020.

[6] S. Fortunato, "Community detection in graphs," *Physics reports*, 2010.

[7] R. Jewkes and A. Murcott, "Meanings of community," *Social science & medicine*, 1996.

[8] B. W. Smith and M. D. Holmes, "Police use of excessive force in minority communities: A test of the minority threat, place, and community accountability hypotheses," *Social problems*, 2014.

[9] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, 1979.

[10] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[11] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *ICML*, 2015.

[12] C. Zhang, Y. Liu, and H. Fu, "Ae2-nets: Autoencoder in autoencoder networks," in *CVPR*, 2019.

[13] H. Wang, L. Zong, B. Liu, Y. Yang, and W. Zhou, "Spectral perturbation meets incomplete multi-view data," *arXiv:1906.00098*, 2019.

[14] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, "Partial multi-view clustering via consistent gan," in *ICDM*, 2018.

[15] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Kloft, and E. Zhu, "Efficient and effective regularized incomplete multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[17] Y. Gong, L. Huang, and L. Chen, "Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method," *arXiv:2101.08533*, 2021.

[18] P. Zeng, M. Yang, Y. Lu, C. Zhang, P. Hu, and X. Peng, "Semantic invariant multi-view clustering with fully incomplete information," *arXiv:2305.12743*, 2023.

[19] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *CVPR*, 2021.

[20] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," *NeurIPS*, 2021.

[21] X. Xiao, B. Wu, J. Wang, C. Li, H. Guo *et al.*, "Seeing the image: Prioritizing visual correlation by contrastive alignment," in *NeurIPS*, 2024.

[22] H. Han, A. J. Wang, P. Ye, and F. Liu, "Unlearning the noisy correspondence makes clip more robust," in *ICCV*, 2025.

[23] Z. Huang, M. Yang, X. Xiao, P. Hu, and X. Peng, "Noise-robust vision-language pre-training with positive-negative learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[24] R. Pan, J. Dong, and H. Yang, "Discovering clone negatives via adaptive contrastive learning for image-text matching," in *ICLR*, 2025.

[25] X. Ma, M. Yang, Y. Li, P. Hu, J. Lv, and X. Peng, "Cross-modal retrieval with noisy correspondence via consistency refining and mining," *IEEE transactions on image processing*, 2024.

[26] Y. Lin, M. Yang, J. Yu, P. Hu, C. Zhang, and X. Peng, "Graph matching with bi-level noisy correspondence," in *ICCV*, 2023.

[27] M. Yang, Z. Huang, and X. Peng, "Robust object re-identification with coupled noisy labels," *International Journal of Computer Vision*, 2024.

[28] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *CVPR*, 2022.

[29] H. Wang, X. Guo, Z.-H. Deng, and Y. Lu, "Rethinking minimal sufficient representation in contrastive learning," in *CVPR*, 2022.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[31] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *NeurIPS*, 2021.

[32] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *CVPR*, 2019.

[33] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama, "Learning discrete representations via information maximizing self-augmented training," in *ICML*, 2017.

[34] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2018.

[35] H. Li, Y. Li, M. Yang, P. Hu, D. Peng, and X. Peng, "Incomplete multi-view clustering via prototype-based imputation," in *IJCAI*, 2023.

[36] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.

[37] M. R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," *NeurIPS*, 2009.

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[39] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *SIGSPATIAL GIS*, 2010.

[40] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[41] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *ACM MM*, 2012.

[42] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[43] Y. Jiang, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "Dm2c: Deep mixed-modal clustering," *NeurIPS*, 2019.

[44] F. Gong, Y. Nie, and H. Xu, "Gromov-wasserstein multi-modal alignment and clustering," in *CIKM*, 2022.

[45] H. Yu, J. Tang, G. Wang, and X. Gao, "A novel multi-view clustering method for unknown mapping relationships between cross-view samples," in *SIGKDD*, 2021.

[46] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, 1987.

[47] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2011.

[48] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *CVPR*, 2017.

[49] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[50] R. Azad, M. Asadi-Aghbolaghi, S. Kasaei, and S. Escalera, "Dynamic 3d hand gesture recognition by learning weighted depth motion maps," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[51] F. Nie, J. Li, X. Li *et al.*, "Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification." in *IJCAI*, 2016.

[52] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *AAAI*, 2017.

[53] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *CVPR*, 2019.

[54] Y. Liu, L. Wang, Y. Bai, C. Qin, Z. Ding, and Y. Fu, "Generative view-correlation adaptation for semi-supervised multi-view learning," in *ECCV*, 2020.

Haobin Li received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2022. He is currently pursuing the Ph.D. degree in computer science at the College of Computer Science, Sichuan University. His research interests include multi-modal learning, noisy correspondence learning, and domain adaptation.

Yijie Lin received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2020. He is currently pursuing the PhD degree with the School of Computer Science, Sichuan University. His research interest includes multi-modal and multi-view learning. On these areas, he has authored more than 10 articles in the top-tier conferences and journals.

Peng Hu received the Ph.D. degree in computer science and technology from Sichuan University, China, in 2019. He is currently an associate research professor at the College of Computer Science, Sichuan University. His research interests mainly focus on multimodal learning, cross-modal retrieval, and network compression. On these areas, he has authored more than 50 articles in the top-tier conferences and journals.

Mouxing Yang is currently pursuing the Ph.D. degree in computer science at the College of Computer Science, Sichuan University. His research interests include multi-modal learning and noisy correspondence learning. On these areas, he has authored 20 articles in the top-tier conferences and journals.

Xi Peng is currently the Cheung Kong distinguished professor at the College of Computer Science, Sichuan University. His current interests mainly focus on machine learning, multimedia analysis, and AI4Science. In these areas, he has co-authored around 100 articles in Nature Communications, JMLR, TPAMI, ICLR, ICML, NeurIPS, etc. Dr. Peng has served as an Associate Editor for five journals including IEEE TPAMI and IEEE TIP.

Supplementary Material of Community-aware Multi-view Representation Learning with Incomplete Information

Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, Xi Peng

In this supplementary material, we provide the mathematical derivation to bridge the relationship between the community versatility learning loss $\mathcal{L}_{cvl}$ and the information theory. Besides, we present more experimental analysis for the proposed CAMERA.

1 Proofs of the Theoretical Analysis

In this section, we provide detailed proofs of Theorem 1 and Theorem 2 and the effect of $\mathcal{L}_{cvl}$ from an informative perspective.

1.1 Proofs of Theorem 1

Theorem 1. *The mutual information between mutual attention A^v and samples Z^v is the lower bound of community commonality $I\left(Z^v; U^v\right)$, i.e., $I\left(Z^v; U^v\right) \geq I\left(Z^v; A^v\right) = -H\left(A^v | Z^v\right) + H\left(A^v\right)$.*

Proof. Since the conditional mutual information $I\left(Z^v; U^v | A^v\right) \geq 0$, the lower bound of the community commonality for the v-th view is

$$\begin{aligned} I\left(Z^v; U^v\right) &= I\left(Z^v; A^v\right) + I\left(Z^v; U^v | A^v\right) - I(Z^v; A^v \mid U^v) \\ &= I\left(Z^v; A^v\right) + I\left(Z^v; U^v | A^v\right) \\ &\geq I\left(Z^v; A^v\right) \\ &= -H\left(A^v | Z^v\right) + H\left(A^v\right), \end{aligned} \tag{1}$$

where $I(Z^v; A^v \mid U^v) = 0$ since we assume that $H(A^v \mid U^v) = 0$, $H(A^v \mid Z^v)$ denotes the conditional entropy of the mutual attention given Z^v, and $H(A^v)$ denotes the entropy of the mutual attention. □

1.2 Proofs of Theorem 2

Theorem 2. *The joint function of community commonality $I\left(Z^v; U^v\right)$ and mutual information between cross-view community centers $I\left(U^1; U^2\right)$ is the lower bound of community versatility, formally,*

$$\begin{aligned} & H\left(U^1 | U^2\right) + H\left(U^2 | U^1\right) \\ & \geq \sum_{v=1}^{2} I\left(Z^v; U^v\right) - 2I\left(U^1; U^2\right). \end{aligned} \tag{2}$$

Proof. According to the the property of mutual information [1], *i.e.*, $H\left(A\right) = I\left(A; B\right) + H\left(A | B\right)$, the community versatility could be represented as,

$$H(U^1 | U^2) + H(U^2 | U^1) = H(U^1) + H(U^2) - 2I(U^1; U^2). \tag{3}$$

Then, the entropy of the community centers is lower bound by the community commonality, *i.e.*,

$$\begin{aligned} \sum_{v=1}^{2} H(U^v) &= \sum_{v=1}^{2} \left(I(Z^v; U^v) + H(U^v | Z^v)\right) \\ &\geq \sum_{v=1}^{2} I(Z^v; U^v) \end{aligned}, \tag{4}$$

where the conditional entropy $H(U^v | Z^v) \geq 0$ holds. According to Eq. 3-4, we derive that the lower bound of community versatility, formally,

$$H(U^1 | U^2) + H(U^2 | U^1) \geq \sum_{v=1}^{2} I(Z^v; U^v) - 2I(U^1; U^2), \tag{5}$$

where $I\left(U^1; U^2\right)$ is the mutual information of cross-view community centers. □

1.3 Effect of $\mathcal{L}_{cvl}$ from the Informative Perspective

In this section, we theoretically prove that the community versatility learning loss $\mathcal{L}_{cvl}$ helps to optimize the mutual information across views (*i.e.*, $I\left(U^1; U^2\right)$) to a margin.

Following DRC [2], we first derive that the cross-view mutual information could be optimized through modeling the conditional probability $p\left(u_i^1 \mid u_i^2\right)$ and marginal proba-

- *The code is available at https://github.com/XLearning-SCU/2025-TPAMI-CAMERA.*

bility $p\left(u_i^1\right)$,

$$\begin{aligned}
& I\left(U^1; U^2\right) \\
= & \mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{p\left(u_i^1, u_i^2\right)}{p\left(u_i^1\right) p\left(u_i^2\right)}\right] \\
= & -\mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{p\left(u_i^1\right)}{p\left(u_i^1 \mid u_i^2\right)}\right] \\
= & -\mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{p\left(u_i^1\right)}{p\left(u_i^1 \mid u_i^2\right)} \mathbb{E}_{u_j^2} \frac{p\left(u_j^2 \mid u_i^1\right)}{p\left(u_j^2\right)}\right] \\
\approx & -\mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{p\left(u_i^1\right)}{p\left(u_i^1 \mid u_i^2\right)} \frac{1}{N_{neg}} \sum_{u_j^2 \in U_{\text{neg}}^2} \frac{p\left(u_j^2 \mid u_i^1\right)}{p\left(u_j^2\right)}\right] \\
= & \mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{\frac{p\left(u_i^1 \mid u_i^2\right)}{p\left(u_i^1\right)}}{\frac{1}{N_{neg}} \sum_{u_j^2 \in U_{\text{neg}}^2} \frac{p\left(u_j^2 \mid u_i^1\right)}{p\left(u_j^2\right)}}\right],
\end{aligned} \tag{6}$$

where U_{neg}^2 is the set of negative community representations and N_{neg} is the number of negative community representations. Following CPC [3], we assume that the optimal value of the function $f\left(u_i^1, u_i^2\right)$ is proportional to $\frac{p\left(u_i^1\right)}{p\left(u_i^1 \mid u_i^2\right)}$,

$$f\left(u_i^1, u_j^2\right) \propto \frac{p\left(u_j^2 \mid u_i^1\right)}{p\left(u_j^2\right)}.$$

As any positive real function f can be used here, we use the popular cosine function in the contrastive loss [4], *i.e.*, $f\left(u_i^1, u_j^2\right) = exp\left(s\left(u_i^1, u_j^2\right)\right)$. We could further derive the mutual information in Eq. 6 to the form of triplet loss,

$$\begin{aligned}
& I\left(U^1; U^2\right) \\
= & \mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{\frac{p\left(u_i^1 \mid u_i^2\right)}{p\left(u_i^1\right)}}{\frac{1}{N_{neg}} \sum_{u_j^2 \in U_{\text{neg}}^2} \frac{p\left(u_j^2 \mid u_i^1\right)}{p\left(u_j^2\right)}}\right] \\
\approx & \mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{\frac{p\left(u_i^1 \mid u_i^2\right)}{p\left(u_i^1\right)}}{\frac{p\left(u_j^2 \mid u_i^1\right)}{p\left(u_j^2\right)}}\right] \\
= & \mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{f\left(u_i^1, u_i^2\right)}{f\left(u_i^1, u_j^2\right)}\right] \\
= & \mathbb{E}_{\left(u^1, u^2\right)} \log \left[\frac{exp\left(s\left(u_i^1, u_i^2\right)\right)}{exp\left(s\left(u_i^1, u_j^2\right)\right)}\right] \\
= & \mathbb{E}_{\left(u^1, u^2\right)}\left(s\left(u_i^1, u_i^2\right) - s\left(u_i^1, u_j^2\right)\right),
\end{aligned} \tag{7}$$

where u_j^2 is the negative community representation of u_i^1 and the second equation holds as we utilize only one negative pair, *i.e.*, $N_{neg} = 1$. Accordingly, we define,

$$D_i^{1,2} = -s\left(u_i^1, u_i^2\right) + s\left(u_i^1, u_j^2\right). \tag{8}$$

Clearly, minimizing $\sum_i^K D_i^{1,2}$ is equal to maximizing the mutual information $I\left(U^1; U^2\right)$.

Drawn the upon theoretical results, we further employ a margin M to $I\left(U^1; U^2\right)$ in Eq. 7 and prove that the triplet loss could be a solution to optimize $I\left(U^1; U^2\right)$ to a margin,

$$\begin{aligned}
& \left[M - I\left(U^1; U^2\right)\right]_+ \\
\approx & \left[M + \frac{1}{K} \sum_i^K D_i^{1,2}\right]_+ \\
\leq & \frac{1}{K} \sum_i^K \left[M + D_i^{1,2}\right]_+ \\
= & \frac{1}{K} \sum_i^K \left[M - s\left(u_i^1, u_i^2\right) + s\left(u_i^1, u_j^2\right)\right]_+,
\end{aligned} \tag{9}$$

where M is the margin. According to Eq. 9, the triplet loss is the upper bound of the target $\left[M - I\left(U^1; U^2\right)\right]_+$. In other words, minimizing the community-level triplet loss helps to implement the constraint in Eq. 9, *i.e.*, $\min\left[M - I\left(U^1; U^2\right)\right]_+$. In our implementation, we impose a regularization to the positive pairs, which optimizes the similarity of the positives to a specific bound β thus preventing learning identical cross-view communities. Besides, the margin is set to $\sigma(\beta)$. Formally, for a given community representation u_i^1 and the corresponding negative community representation u_j^2, we propose the following loss function,

$$\ell = \left[\sigma(\beta) - \min\left(s\left(u_i^1, u_i^2\right), \beta\right) + s\left(u_i^1, u_j^2\right)\right]_+. \tag{10}$$

2 More in-depth Discussions with the sample–sample Restoration Paradigm.

In this section, we provide detailed discussions about the differences between the existing sample-sample restoration paradigm and the proposed sample-community restoration paradigm. Specifically, as mentioned in Introduction, community commonality and community versatility are indispensable in multi-view representation learning with incomplete information. However, the existing sample-sample restoration paradigm struggles to capture both of them. In the following, we will elaborate on the limitations of the existing sample-sample imputation and alignment paradigms.

The sample-sample imputation paradigm aims to impute the missing samples by leveraging the observed samples. According to the differences in the imputation strategies, the existing methods could be further divided into neighborhood-based methods [14], [15] and generative methods [13], [12]. Specifically, neighborhood-based recovery aims to leverage cross-view neighbors to impute missing samples. Such a paradigm implicitly assumes that the views could be mapped into a common space wherein the neighbors of the missing sample could be accurately identified by its cross-view counterpart. In practice, however, such an assumption is always satisfied at the cost of the community versatility, as the view-specific information is often excluded to learn the common space. To compensate for community versatility, some studies propose capturing the view-specific information using a cross-view predictor [13] or generator [12]. Unfortunately, such a generative paradigm essentially learns an equivalent mapping for the whole dataset across views, which will lose the community commonality.

As for the sample-sample alignment paradigm, it aims to establish cross-view correspondence by leveraging the

TABLE 1
The architecture of the encoders and decoders in CAMERA. In the table, $\dim^{(v)}$ denotes the dimension of input data in the v-th view, $\mathbf{k}$ denotes the kernel size, $\mathbf{s}$ denotes the stride, $\mathbf{p}$ denotes the padding operation and $\mathbf{op}$ denotes the output padding operation.

Dataset	Encoder	Decoder
Scene-15 Reuters LandUse-21 CUB DHA UWA	Linear(dim$^{(v)}$, 1024), BatchNorm, ReLU Linear(1024, 1024), BatchNorm, ReLU Linear(1024, 512), BatchNorm, ReLU Linear(512, 256), ReLU Linear(256, 256)	Linear(256, 256), BatchNorm, ReLU Linear(256, 512), BatchNorm, ReLU Linear(512, 1024), BatchNorm, ReLU Linear(1024, 1024), BatchNorm, ReLU Linear(1024, dim$^{(v)}$)

NoisyMNIST	Resize(28, 28)
	Conv2d(1, 16, **k**=(3, 3), **s**=(1, 1), **p**=(1, 1)), ReLU
	Conv2d(16, 32, **k**=(3, 3), **s**=(2, 2), **p**=(1, 1)), ReLU
	Conv2d(32, 32, **k**=(3, 3), **s**=(1, 1), **p**=(1, 1)), ReLU
	Conv2d(32, 16, **k**=(3, 3), **s**=(2, 2), **p**=(1, 1)), ReLU
	Flatten
	Linear(784, 256), ReLU
	Linear(256, 256)
		Linear(256, 256), ReLU
		Linear(256, 784), ReLU
		Resize(28, 28)
		ConvTranspose2d(16, 32, **k**=(3, 3), **s**=(2, 2), **p**=(1, 1), **op**=(1, 1))
		ConvTranspose2d(32 32, **k**=(3, 3), **s**=(2, 2), **p**=(1, 1))
		ConvTranspose2d(32, 16, **k**=(3, 3), **s**=(2, 2), **p**=(1, 1), **op**=(1, 1))
		ConvTranspose2d(16, 1, **k**=(3, 3), **s**=(2, 2), **p**=(1, 1))
		Flatten

TABLE 2
The multi-view human action recognition performance on the DHA dataset. The best and second best results are denoted in **bold** and underline.

Method	RGB	R → D	Depth	D → R	R + D
LIBSVM [5]	66.1	70.2	78.9	78.2	83.5
VLAD [6]	67.1	-	-	-	-
TSN [7]	67.9	-	-	-	-
WDMM [8]	-	-	**81.1**	-	-
AMGL [9]	64.6	59.1	72.8	67.3	74.9
MLAN [10]	67.9	67.9	73.0	72.8	76.1
GMVAR [11]	-	69.7	-	**83.5**	88.7
GVCA [12]	-	-	-	-	89.3
DCP [13]	78.4	79.5	79.3	81.0	89.3
Ours	**79.7**	**79.8**	79.8	79.8	**89.9**

relationship between cross-view samples, which could be divided into the following two categories: i) graph matching methods [16], which employ graph matching algorithms such as the Hungarian algorithm to build the cross-view correspondence; ii) instance identification methods [17], which identify the cross-view within-category counterparts for the given samples and re-establish the correspondence between them. However, both graph matching and instance identification methods assume that the views could be mapped into a common space, so that cross-view graph structures and within-category counterparts could be determined. As discussed above, such an assumption is always satisfied at the cost of the community versatility, which is indispensable in MvRL.

Different from the aforementioned methods, the proposed sample-community restoration framework could simultaneously embrace both community commonality and community versatility. On the one hand, with the sample-to-community relationship, CAMERA integrates the corresponding community center into the sample, thus enhancing community commonality. On the other hand, with the community-to-sample relationship, CAMERA aggregates within-community samples to represent the community and then learns view-specific communities, thus preserving community versatility. Thanks to the proposed sample-community restoration framework, the incomplete information could be recovered. Specifically, CAMERA employs the sample-community relationship of the observed samples and community centers of the missing view to impute the missing samples, while establishing the correspondences by identifying the sample-community relationships of cross-view samples.

3 MULTI-VIEW DATASETS

In this section, we present the details of the datasets used in the manuscript as follows.

- **Scene-15** [18]: Scene-15 comprises a collection of 4,485 images captured from 15 indoor and outdoor scenes/categories. Following [17], we leverage the 20-dimensional GIST features and the 59-dimensional PHOG features as two distinct views.
- **Reuters** [19]: Reuters is a multilingual news dataset consisting of 18,758 instances across six languages. Following [16], we project the two selected languages, namely, English and French, into a 10-dimensional latent space through PCA [20].
- **NoisyMNIST** [21]: NoisyMNIST is a large-scale multi-view dataset that is composed of 70,000 instances distributed across 10 categories. We select a subset of 30,000 instances for evaluation since some of the baselines cannot handle the original scale of the dataset.
- **CUB** [22]: CUB consists of 11,788 images and their corresponding captions, representing 200 subcategories of birds. Following [23], we extract deep visual features through the GoogLeNet, while texture features are obtained using the doc2vec approach. In our experiment, we focus on the first 10 categories derived from these two views.
- **LandUse-21** [24]: LandUse-21 contains 2,100 satellite images of 21 categories, and we utilize the 59-dimensional PHOG features and the 40-dimensional LBP features as the two views.
- **UWA** [25]: The dataset consists of 660 action sequences instances of 30 categories. The two views are the 6144-dimensional RGB features and the 110-dimensional depth features.
- **DHA** [26]: The dataset contains 483 video clips of 23 categories with the 6144-dimensional RGB features and the 110-dimensional depth features.

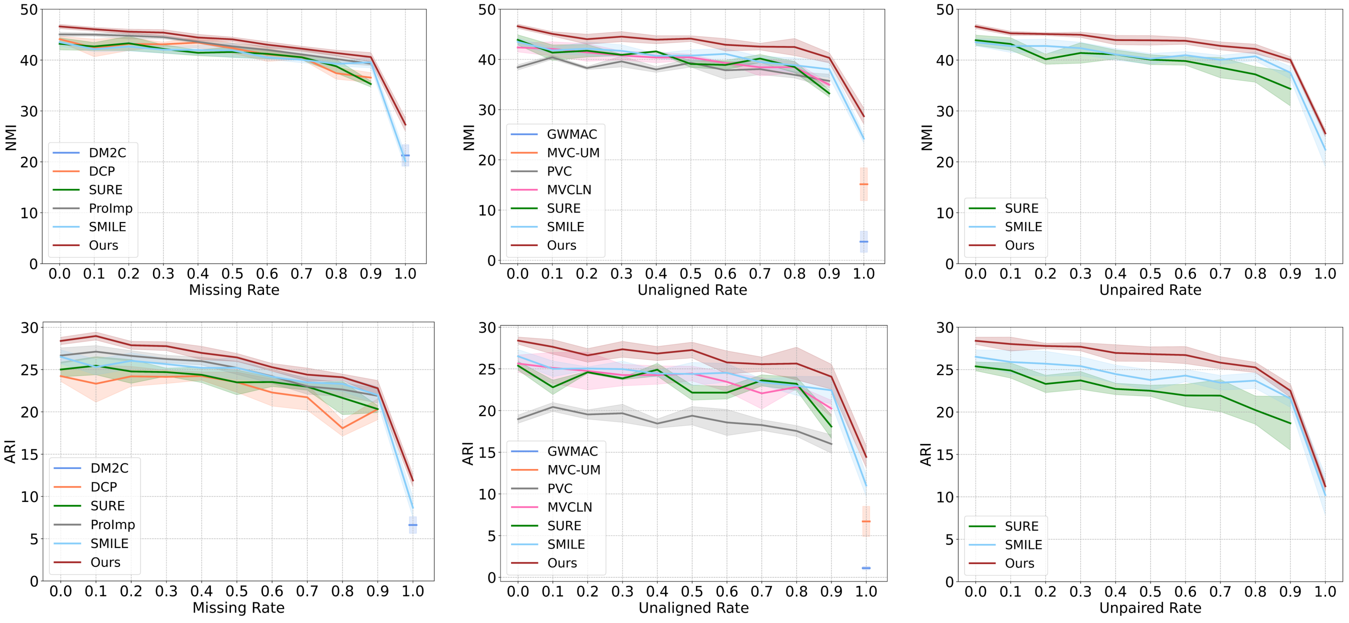
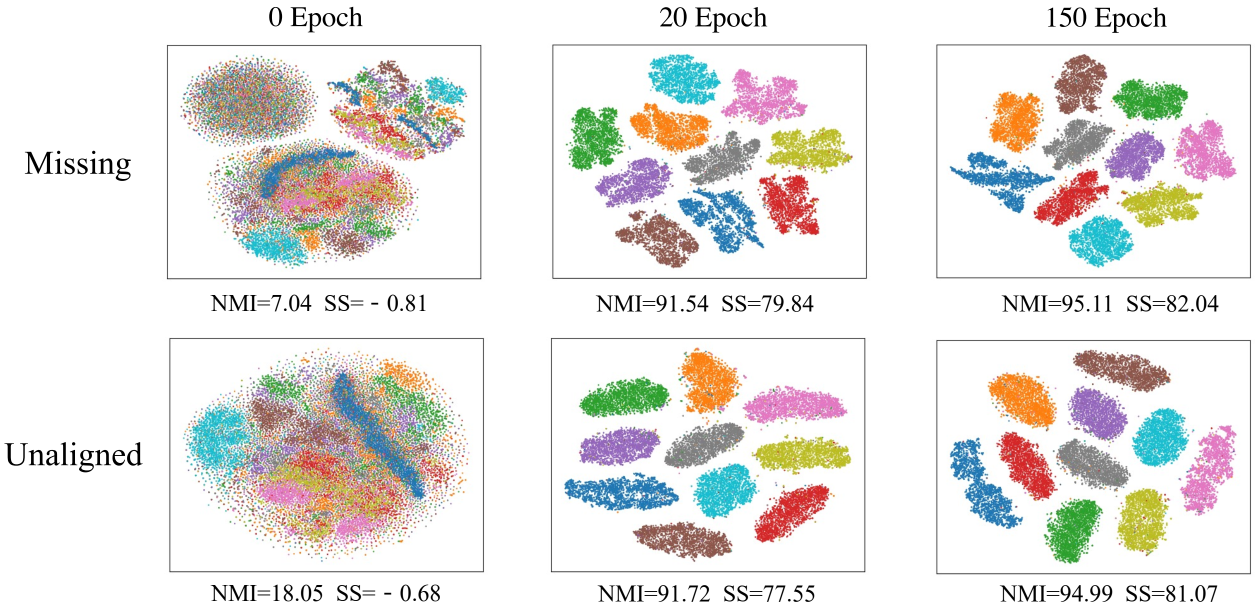


Fig. 1. Multi-view clustering performance on the Scene-15 dataset with different Missing/Unaligned/Unpaired rates in terms of NMI and ARI.



NMI=7.04 SS= - 0.81

NMI=91.54 SS=79.84

NMI=95.11 SS=82.04

NMI=18.05 SS= - 0.68

NMI=91.72 SS=77.55

NMI=94.99 SS=81.07

Fig. 2. t-SNE visualization on the NoisyMNIST dataset across the training process. In the figure, SS denotes the silhouette score.

4 Network Architectures of CAMERA

The proposed CAMERA employs auto-encoders architecture whose details are presented in Tab. 1. Specifically, following [27], [13], we employ the convolutional auto-encoder for the multi-view image dataset (NoisyMNIST) and adopt the full-connected auto-encoder for the other datasets.

5 More Detailed Experiments

In this section, we provide more experiment results to further verify the effectiveness of the proposed CAMERA.

5.1 Additional Experiment Results on Human Recognition Task

To prove the effectiveness of CAMERA in the human action recognition task, we have compared CAEMRA with the state-of-the-art methods on the UWA dataset in Section 4.5 of the manuscript. To further verify the generality of CAMERA, we additional carry out experiments on the DHA dataset with the same setting as the manuscript. From the results in Tab. 2, one could see that CAMERA significantly improves the performance of most settings.

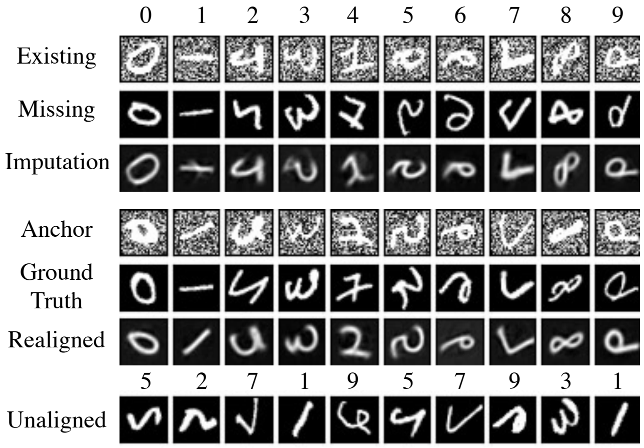


Fig. 3. Data restoration on the NoisyMNIST dataset. We show the data restoration results of “Missing” setting and “Unaligned” setting in the Row 1-3 and Row 4-7, respectively. For the “Missing” setting, the rows represent the existing view, the missing view, and the imputed view. For the “Unaligned” setting, the rows correspond to the anchor view, the view with ground truth correspondence, the aligned samples, and the unaligned view.

5.2 More Analytical Experiments

In the manuscript, we have verified the robustness of CAMERA on various Missing/Unaligned/Unpaired rates (Section 4.3). Here, we provide the experiments on the other two evaluation metrics, *i.e.*, NMI and ARI. As shown in Fig. 1, CAMERA outperforms all the baselines under all settings, which illustrates the robustness of CAMERA against incomplete information.

5.3 t-SNE Visualization

We conduct t-SNE visualization on the fusion representation at distinct training epochs. From the result in Fig. 2, one could observe that: i) at the 0 epoch, the imputed and aligned representations are collapsed, since the community centers are randomly initialized and the mutual attention is incorrect. In the “Missing” setting, data forms three clusters at initialization, which corresponds to a group of complete data and two groups of imputed data (missing in different views). In the “Unaligned” setting, the data forms one cluster, which corresponds to the group of unpaired data; ii) as the epoch increases, thanks to the community commonality induced by our CAMERA, the imputed and aligned data forms more compact clusters, thus boosting clustering performance.

5.4 Visualization on Data Restoration

To further verify the restoration ability of CAMERA, we perform visualization analysis in the raw data space by decoding the imputed and aligned representation. In the experiments, we impute the missing samples based on the observed view and establish the correspondence based on the anchor view. From the results in Fig. 3, CAMERA successfully imputes the missing samples and establishes the correspondence. It is worth noting that the imputed and aligned samples share the same category information as the observed samples, which illustrates the effectiveness of the MA-based restoration framework.

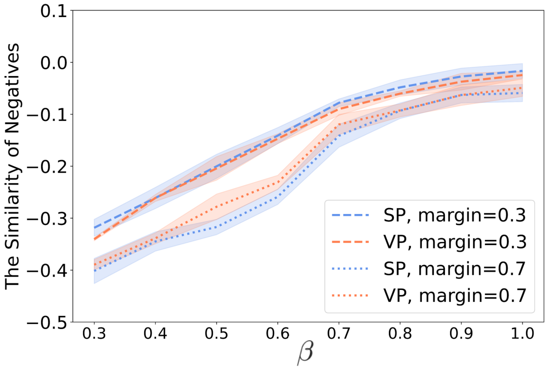


Fig. 4. The parameter analysis of the margin.

5.5 Influence of the Margin in $\mathcal{L}_{cvl}$

In the proposed community versatility learning loss, we implement the margin as $\sigma(\beta) = \beta^2$ to prevent the similarity of negatives from wrongly increasing as β increases. To prove the effectiveness of the proposed strategy, we set the margin to 0.3 and 0.7 and then conduct the training process. As shown in Fig. 4, simply setting the margin to a fixed parameter would significantly increase the similarity of negatives as the bound β increases.

References

[1] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[2] H. Zhong, C. Chen, Z. Jin, and X.-S. Hua, “Deep robust clustering by contrastive learning,” *arXiv:2008.03030*, 2020.

[3] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv:1807.03748*, 2018.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML*, 2020.

[5] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, 2011.

[6] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, “Actionvlad: Learning spatio-temporal aggregation for action classification,” in *CVPR*, 2017.

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *ECCV*, 2016.

[8] R. Azad, M. Asadi-Aghbolaghi, S. Kasaei, and S. Escalera, “Dynamic 3d hand gesture recognition by learning weighted depth motion maps,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[9] F. Nie, J. Li, X. Li *et al.*, “Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification.” in *IJCAI*, 2016.

[10] F. Nie, G. Cai, and X. Li, “Multi-view clustering and semi-supervised classification with adaptive neighbours,” in *AAAI*, 2017.

[11] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, “Generative multi-view human action recognition,” in *CVPR*, 2019.

[12] Y. Liu, L. Wang, Y. Bai, C. Qin, Z. Ding, and Y. Fu, “Generative view-correlation adaptation for semi-supervised multi-view learning,” in *ECCV*, 2020.

[13] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, “Dual contrastive prediction for incomplete multi-view representation learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] H. Tang and Y. Liu, “Deep safe incomplete multi-view clustering: Theorem and algorithm,” in *ICML*, 2022.

[15] G. Chao, Y. Jiang, and D. Chu, “Incomplete contrastive multi-view clustering with high-confidence guiding,” in *Proceedings of the AAAI conference on artificial intelligence*, 2024.

[16] Z. Huang, P. Hu, J. T. Zhou, J. Lv, and X. Peng, "Partially view-aligned clustering," *NeurIPS*, 2020.

[17] M. Yang, Y. Li, P. Hu, J. Bai, J. C. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[18] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.

[19] M. R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," *NeurIPS*, 2009.

[20] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, 2010.

[21] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *ICML*, 2015.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[23] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *SIGSPATIAL GIS*, 2010.

[25] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[26] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *ACM MM*, 2012.

[27] P. Zeng, M. Yang, Y. Lu, C. Zhang, P. Hu, and X. Peng, "Semantic invariant multi-view clustering with fully incomplete information," *arXiv:2305.12743*, 2023.